BRIEF CONTENTS

PART ONE: Principles of Test Construction

1. Introduction to Measurement **3**
2. Test Construction **21**
3. Item Writing **36**
4. Item Analysis **57**
5. Reliability **79**
6. Validity **104**
7. Norms and Test Scales **125**
8. Response Set in Test Scores **140**

PART TWO: Principles of Measurement

9. Measurement of Intelligence, Aptitude and Achievement **149**
10. Measurement of Personality **186**
11. Projective Techniques **219**
12. Techniques of Observation and Data Collection **267**
13. Scaling Techniques **315**

PART THREE: Principles of Research Methodology

14. Sampling **365**
15. Social Scientific Research **392**
16. Single-Subject Experimental Research and Small *N* Research **433**
17. Historical Research **446**
18. The Problem and The Hypothesis **451**
19. Reviewing the Literature **466**
20. Variables **474**
21. Research Design **491**
22. Qualitative Research **555**
23. Carrying Out Statistical Analyses **593**
24. Writing a Research Report and a Research Proposal **669**

Objective Questions 683

Appendices and References 711

Glossary 753

Subject Index 777

SHOUT
SHARE
IT

CONTENTS

PART ONE: Principles of Test Construction

1. **INTRODUCTION TO MEASUREMENT 3**
    Measurement and Evaluation **3**
    The History of Psychological Measurement and Mental Testing **4**
    Levels of Measurement (or Measurement Scales) **9**
    Properties of Scales of Measurement **13**
    Functions of Measurement **13**
    Distinction between Psychological Measurement and Physical Measurement **15**
    Steps in Measurement Process **16**
    Problems Related to the Measurement Process **17**
    General Problems of Measurement **18**
    Sources of Errors in Measurement **19**
    Difference among Assessment, Testing and Measurement **20**
    Review Questions **20**

2. **TEST CONSTRUCTION 21**
    Meaning of Test in Psychology and Education **21**
    Classification of Tests **23**
    Characteristics of a Good Test **25**
    General Steps of Test Construction **26**
    Uses and limitations of Psychological Tests and Testings **29**
    Ethical Issues in Psychological Testing **31**
    Steps in Selecting Appropriate Published Tests **33**
    Review Questions **35**

3. **ITEM WRITING 36**
    Meaning and Types of Items **36**
    Difference Between Essay-Type Tests and Objective-Type Tests **51**
    General Guidelines for Item Writing **54**
    General Methods of Scoring Objective-Test Items **55**
    Review Questions **56**

4. **ITEM ANALYSIS 57**
    Meaning and Purpose of Item Analysis **57**
    Power Tests **58**
    Item Difficulty **58**
    Optimal Difficulty Value for a Reliable Test **63**
    Index of Discrimination **63**
    A Simplified Item Analysis Form **67**
    Effectiveness of Distractors or Foils (Distractor Analysis) **69**
    Speed Tests **70**
    Factors Influencing the Index of Difficulty and the Index of Discrimination **72**
    Problems of Item Analysis **72**

Important Interactions Among Item characteristics **74**
The Item Characteristic Curve (ICC) and Item Response theory **75**
Review Questions **78**

5. **RELIABILITY 79**
History and Theory of Reliability **79**
Meaning of Reliability **82**
Methods (or Types) of Reliability **85**
What is a Satisfactory Size for the Reliability Coefficient? **95**
Standard Error of Measurement **96**
Reliability of Speed Test **97**
Factors Influencing Reliability of Test Scores **97**
How to Improve Reliability of Test Scores? **100**
Estimation of True Scores **101**
Index of Reliability **102**
Reliability of Difference Score **102**
Reliability of Composite Score **103**
Review Questions **103**

6. **VALIDITY 104**
Meaning of Validity **104**
Aspects of Validity **105**
Convergent Validation and Discriminant Validation **114**
Statistical Methods for Calculating Validity **117**
Factors Influencing Validity **118**
Concept of Cross-Validation **121**
Extravalidity Concerns **122**
Relation of Validity to Reliability **123**
Review Questions **124**

7. **NORMS AND TEST SCALES 125**
Meaning of Norm-Referencing and Criterion-Referencing **125**
Steps in Developing Norms **127**
Types of Norms and Test Scales **127**
Computer Applications in Psychological Testing and Assessment **136**
Criteria of Good Test Norms **138**
Review Questions **139**

8. **RESPONSE SET IN TEST SCORES 140**
Meaning of Response Sets **140**
Types of Response Sets **141**
Implications of Response Sets **143**
Methods to Eliminate Response Sets **143**
Review Questions **146**

PART TWO: Principles of Measurement

9. **MEASUREMENT OF INTELLIGENCE, APTITUDE AND ACHIEVEMENT 149**
Different Viewpoints Towards Intelligence **149**
Types of Intelligence Tests **157**

Types of Intelligence Test Scores 175
Distinction Between Aptitude Test and Achievement Test 177
Types of Aptitude and Achievement Tests 178
Tests of Creativity 184
Review Questions 185

10. MEASUREMENT OF PERSONALITY 186

Meaning and Purpose of Personality Measurement 186
Tools of Personality Assessment 187
Popular Strategies Involved in Construction of Personality Inventories 188
Overcoming Distortions in Self-report Inventories 206
Situational Tests 207
Measurement of Interests, Values and Attitudes 209
Review Questions 218

11. PROJECTIVE TECHNIQUES 219

Meaning and Types of Projective Techniques 219
Classification of Projective Techniques 220
Pictorial Techniques 224
The Rorschach Test 224
Interpretation of the Rorschach Protocol 241
The Holtzman Inkblot Test 252
Thematic Apperception Test (TAT) 254
Verbal Techniques 260
Expressive Techniques 261
Evaluation of Projective Techniques 264
Review Questions 266

12. TECHNIQUES OF OBSERVATION AND DATA COLLECTION 267

Methods of Data Collection 268
Questionnaire and Schedule 269
Interview 275
Content Analysis 280
Observation as a Tool of Data Collection 282
Difference Between Participant Observation and Nonparticipant Observation 285
Rating Scale 285
Types of Rating Scales 286
Other Special Types of Rating Scales 293
Problems in Obtaining Effective Ratings 305
Methods of Improving Effectiveness of Rating Scales 307
Errors in Ratings 308
Evaluation of Rating Scales 310
Review Questions 313

13. SCALING TECHNIQUES 315

Distinction Between Psychophysical and Psychological Scaling Methods 315
Psychophysical Scaling Methods 316
Weber's Law and Fechner's Law 333
Steven's Power Law 335

Newer Psychophysical Methods **336**
Psychological Scaling Methods **337**
Methods of Attitude Scales or Opinionnaires **352**
Review Questions **362**

PART THREE: Principles of Research Methodology

14. SAMPLING 365

Meaning and Types of Sampling **365**
Need for Sampling **367**
Fundamentals of Sampling **368**
Principles of Sampling **369**
Factors Influencing Decision to Sample **371**
How Large should a Sample Be? **372**
Methods of Drawing Random Samples **374**
Simple Random Sample **375**
Stratified Random Sample **378**
Area (or Cluster) Sampling **381**
Quota Sampling **382**
Purposive or Judgemental Sampling **383**
Accidental Sampling **384**
Snowball Sampling **384**
Saturation Sampling and Dense Sampling **385**
Double Sampling **386**
Mixed Sampling **386**
Requisites of a Good Sampling Method **387**
Common Advantages of Sampling Methods **388**
Sampling Distribution **388**
Sampling Error **389**
Review Questions **391**

15. SOCIAL SCIENTIFIC RESEARCH 392

Meaning and Characteristics of Scientific Research **392**
Scientific Approach to the Study of Behaviour **394**
Validity in Research or Experimental Validity **397**
Controlling Threats to Reliability and Validity in Research **401**
Phases or Stages in Research **403**
Types of Educational Research **406**
Types of Research: Experimental and Nonexperimental **406**
Difference Between Research Method and Research Methodology **424**
Ethical Problems in Research **425**
Comparison Between Experimental and Nonexperimental Research **427**
Types of Experiment **428**
Types of Applied Research **429**
Review Questions **431**

16. SINGLE-SUBJECT EXPERIMENTAL RESEARCH AND SMALL *N* RESEARCH 433

Meaning and Origin of Single-Subject Experimental Research **433**
General Procedures of Single-Subject Experimental Research **434**

Basic Designs of Single-Subject Experimental Research 435
Data-Collection Strategies in Single-Subject Experimental Research 439
Evaluating Data in Single-Subject Experimental Research 440
Strengths and Weaknesses of Single-Subject Experimental Research 441
Comparison Between Single-Subject Research and Large *N* Research 443
Small *N* Design: Nature and Historical Perspectives 444
Review Questions 445

17. **HISTORICAL RESEARCH 446**
Meaning of Historical Research and its Necessity 446
Steps in Historical Research 447
Sources of Historical Data 447
Historical Criticism 448
Limitations of Historical Research 449
Review Questions 450

18. **THE PROBLEM AND THE HYPOTHESIS 451**
Meaning and Characteristics of a Research Problem 451
Sources of Stating a Research Problem 453
Important Considerations in Selecting a Research Problem 454
Ways in Which a Problem is Manifested 455
Types of Research Problems 456
Importance of Formulating a Research Problem 457
Steps in Formulating a Research Problem 457
Meaning and Characteristics of a Good Hypothesis 458
Formulating a Hypothesis 460
Ways of Stating a Hypothesis 461
Types of Hypotheses 461
Sources of Hypotheses 463
Functions of Hypotheses 464
Review Questions 465

19. **REVIEWING THE LITERATURE 466**
Purpose of the Review 466
Types of Literature Review 467
Sources of the Review 467
Types of Literature 469
Writing Process of the Literature Review 470
How Old should the Literature be? 471
Preparation of Index Card for Reviewing and Abstracting 471
Abstract 472
Review Questions 473

20. **VARIABLES 474**
Meaning and Types of Variables 474
Difference between a Variable and a Concept 481
Methods of Measuring Dependent Variables 481
Important Considerations in Selection of Variables 482
Important Approaches to Manipulating Independent Variables 483
Techniques of Controlling Extraneous Variables 484
Controlling Demand Characteristics 487
Review Questions 490

21. RESEARCH DESIGN 491
- Meaning and Purpose of Research Design 492
- Criteria of Research Design 494
- Basic Principles of Experimental Design 495
- Basic Terms used in Experimental Design 496
- Some Important Types of Research Design 500
- Between-subjects Design 501
- Problem of Creating Equivalent Groups in Between-subjects Design 534
- Within-subjects Design 535
- Problem of Controlling Sequence Effects in Within-subjects Design 536
- Comparison of Between-subjects Design and Within-subjects Design 538
- Experimental Design based upon the Campbell and Stanley Classification 539
- Pre-Experimental Design (Nondesigns) 540
- True Experimental Design 541
- Quasi-Experimental Designs 543
- Ex Post Facto Design 550
- Steps in Experimentation 551
- Review Questions 554

22. QUALITATIVE RESEARCH 555
- Meaning and Essential Features of Qualitative Research 555
- A Qualitative Research Model: Five Components 556
- Relevance of Qualitative Research 558
- Qualitative Research: A Brief Historical Introduction 558
- Themes of Qualitative Research 560
- Theoretical Perspectives of Qualitative Research 562
- Research Design Strategies of Qualitative Research 564
- Sampling Techniques of Qualitative Research 569
- Data Collection Techniques in Qualitative Research 572
- Data Analysis and Interpretation 578
- Comparison of Methods of Qualitative and Quantitative Data Analysis 587
- Combining Qualitative and Quantitative Approaches 589
- Review Questions 591

23. CARRYING OUT STATISTICAL ANALYSES 593
- Sample and Population 593
- Normal Curve 594
- Measures of Relative Position: Standard Scores 600
- Parametric and Nonparametric Statistical Tests 601
- Parametric Statistics 604
- Nonparametric Statistics 641
- Correlation and Regression 659
- Major Terms and Issue in Correlation and Regression 660
- Choosing Appropriate Statistical Tests 664
- Review Questions 668

24. WRITING A RESEARCH REPORT AND A RESEARCH PROPOSAL 669
- General Purpose of Writing a Research Report 669

Structure or Format of a Research Report (Style Manual) 669
Style of Writing a Research Report 674
Typing the Research Report 676
Evaluating a Research Report 676
Preparing a Research Proposal 676
Review Questions 679

OBJECTIVE QUESTIONS 683

APPENDICES AND REFERENCES 711

GLOSSARY 753

SUBJECT INDEX 777

◆

Part One

PRINCIPLES OF TEST CONSTRUCTION

1

INTRODUCTION TO MEASUREMENT

CHAPTER PREVIEW

- Measurement and Evaluation
- The History of Psychological Measurement and Mental Testing
- Levels of Measurement (or Measurement Scales)
  - (i) Nominal or Classificatory Measurement
  - (ii) Ordinal Measurement
  - (iii) Interval or Equal Interval Measurement
  - (iv) Ratio Measurement
- Properties of Scales of Measurement
- Functions of Measurement
- Distinction between Psychological Measurement and Physical Measurement
- Steps in Measurement Process
- Problems Related to Measurement Process
- General Problems of Measurement
  - (i) Indirectness of Measurement
  - (ii) Incompleteness of Measurement
  - (iii) Relativity of Measurement
  - (iv) Errors in Measurement
- Sources of Errors in Measurement
- Difference among Assessment, Testing and Measurement

MEASUREMENT AND EVALUATION

In everyday life, the terms 'measurement' and 'evaluation' are often used interchangeably. However, in psychological, sociological and educational researches these two terms are used separately because they connote two different meanings. 'Measurement' refers to the process of assigning numerals to events, objects, etc., according to certain rules. Tyler (1963, 7) defines measurement as "assignment of numerals, according to rules". A still more elaborate and wider definition has been given by Nunnally (1970, 7): "Measurement consists of rules for assigning numbers to objects in such a way as to represent quantities of attributes." An analysis of the definition of measurement given by Nunnally reveals the following main properties of measurement.

1. In the process of measurement, numbers are assigned according to some rules. A number is a kind of numeral to which is assigned some quantitative meaning. In the process of measurement the investigator assigns numbers, not of his own choice, but according to certain fixed and explicit rules. Rules are the procedures to transform qualities of attributes into numbers (Nunnally & Bernstein 1994; Rosenthal, Rasnow & Rubin 2000). Usually, such rules are of different types. One type is where the procedure is obvious and explicit. For example, when one is measuring the length of cloth in feet and inches, rules for assigning numerals are very explicit and clear. But suppose one were to measure the extroversion trait of personality or the

intelligence of a child. Obviously, in such a situation the rules would not be as clear as in the first example. For measuring psychological, sociological and educational attributes the rules are generally vague and less explicit.

2. Measurement is always concerned with certain attributes or features of the object. It is these attributes or features of the object which are measured and not the object itself. For example, one would measure the aptitude, intelligence, attitude, etc., of a person and not the person himself. When an investigator is measuring the attribute of a person, he is faced with two difficulties. *First*, he may be asked to measure an attribute, the existence of which is doubtful. Extra-sensory perception is one such example. Most investigators have failed to show such perception in many individuals. In such a case, measurement is not difficult, rather it is an impossible task. *Second*, the investigator may be asked to measure attributes which are not unitary, rather they are a mixture of several subattributes. Usually, this happens when one is asked to measure 'personality adjustment'. Adjustment may relate to home, school, emotion, etc., each of which requires a separate measurement and any attempt to measure them together may create difficulty. In such a situation the investigator can, however, measure the attributes with more sophisticated instruments specially designed for the purpose.

3. In the process of measurement, numerals are used to represent quantities of the attribute. In other words, measurement involves the process of quantification. Quantification indicates how much or to what extent that particular attribute is present in a particular object. For example, when the investigator is measuring the achievement of a child in arithmetic, he quantifies it by saying that the child has 80% marks in his class. The percentage indicates how much of arithmetical knowledge he has gained in the class.

Measurement is different from its so-called synonym 'evaluation'. By evaluation, we mean appraisal or assessment with respect to some standard. Tuckman (1975, 12) defines evaluation as "a process wherein the parts, processes or outcomes of a programme are examined to see whether they are satisfactory, particularly with reference to the programme's stated objectives, our own expectations or our own standards of excellence." Thus evaluation involves a process of appraisal of an object or event with reference to some standard. The standard may be social, cultural or scientific. The standard may also be true or arbitrary. An investigator may measure the height of a 6-year-old child (which is, say, 30") and type him as short. A typist typing 80 words per minute may be described as a 'Grade A' typist. Description of the height of the child (which is 30") and the typing speed of the typist (which is 80 words per minute), are examples of measurement. However, when the child is said to be short or the typist is classified as a 'Grade A' typist, it means the performance of the typist and the height of a child are being compared with reference to some standard. A child is short because he is shorter than the general mean height of children of his age group and the typist is a 'Grade A' typist because his speed is faster than the average speed of most typists. Thus, the height of the child and the typing behaviour of the typist are being evaluated (and are not being measured).

The History of Psychological Measurement and Mental Testing

Measurement is the root of both biological and behavioural sciences. In biological science, measurement has a long history but the history of psychological and educational measurement dates back to around 206 BC when Chinese officials were examined every third year to determine their fitness for office (Gregory 2005). During the Han Dynasty period (206 BC to 220 AD), the use of test batteries became very common. Tests had become an established and accepted instrument for such diverse topics like civil law, military affairs, revenue, agriculture, etc. During the Ming Dynasty (1368–1644 AD), tests for these various services were used as an accepted norm. During this period, a national multistage testing programme involved local and regional testing centres. Those who passed the tests at the local level, went on to provincial capitals for extensive essay examination and those who passed the provincial examinations went on to the nation's capital for a final round. Only those who passed this third set of tests were declared fit for public offices.

The western countries learnt about the testing programmes through the Chinese's efforts. The British government encouraged the East India Company in 1832 to follow the Chinese system as a method of selecting employees for overseas duty. Even the British government itself adopted a similar system of testing for its civil services in 1855. After this, the French and the German governments also followed suit. In 1883, the US government established the American Civil Service Commission that conducted various competitive examinations for certain government jobs. The impact of such testing programmes grew rapidly in other western countries, too.

Most historians today fully agree that the beginning of psychological testing relate to the experimental investigation of individual differences that flourished in Germany and Great Britain after 1850. Prior to 1850, psychology was not considered to be an independent science but only a branch of philosophy. Any attempt to measure human behaviour through experiments was ridiculed by philosophers. According to Kant, measuring human behaviour was useless, because human reactions were such that they could not be observed and measured directly. But, beginning with 1850, psychology began to shake off this old relation with philosophy and started forming a new alliance with biological sciences. The concepts that were discovered in the fields of physics, chemistry and biology were proving to be of immense help in measuring human characteristics such as forgetfulness, intelligence, response time, sensation, perception, etc. Biologists and physicists were pioneers in conducting experiments for measuring human behaviour. Influenced by their experiments, psychology started getting measurement-conscious and the movement of measurement was being spread through experimentation and demonstrations. As a matter of fact, the history of psychological and educational measurement during the period 1850 to 1900 was nourished by three major influences, namely, psychophysics, Darwinian biology and clinical practices.

The nineteenth century, particularly the second half, had witnessed the development of a number of scientific tools and innovations in physics and chemistry, and scientists were of the opinion that those physical methods could also be successfully tried in measuring psychological properties. As a result, psychophysics was born and experimental psychologists began to study the relationship between physical stimuli such as the intensity of light or sound, and the experiences or sensations produced by these stimuli. Many new experimental techniques of measurement followed these psychophysical experimentations. In 1885, the German physician Hubert Von Grashey developed what is called antecedent of the memory drum as a means of testing patients with brain injury. His report revealed that many patients could clearly recognize stimuli but could not identify them when displayed through the moving slot. Likewise, German psychiatrist Conrad Rieger developed another test battery for the brain damaged. The demerit of this test was that it took over 100 hours in its administration and therefore, soon went out of use.

Besides psychophysical methods, some early physiologists were also influencing the development of psychological measurement. These physiologists were mainly interested in measuring the processes of seeing, hearing, speed of conduction, etc. As a result, the first laboratory for studying psychological reactions was established in 1879 by Wilhelm Wundt at the University of Leipzig (now Karl Marx University). Although the laboratory was established for measuring and studying psychological reactions, Wundt and his disciples mostly concentrated upon physiological reactions only. Gradually, however, this laboratory flourished and developed into an excellent centre for the measurement of human reactions and invited the attention of several experimenters from abroad, too. Although Wundt is credited with founding the first psychological laboratory in 1879 in Leipzig, very few know that he was measuring mental processes for years, at least as early as 1862. He conducted an experiment with his thought meter, which was a calibrated pendulum with needle swinging back and forth and striking the bells. The purpose of this thought meter was to assess the swiftness of the thought of the testee. Wundt's claim was that the difference between the observed pendulum position and actual position would easily provide a way for determining the swiftness of thought of the testee. This analysis by Wundt's was relevant for studying astronomer's longstanding problem. The problem was that when two or more astronomers simultaneously use the same telescope for noting the crossing times as the stars moved across a grid line on a telescope, there existed a difference. Even in Wundt's time, a remarkable event occurred in the history of science—Kinnebrook, an assistant at the Royal Observatory in England, was dismissed in 1796 simply because his assessment of stellar crossing times was about one second slower (Boring 1950).

Darwin's theory of evolution also provided a big impetus to the growth of psychological measurement. Although it originated in biology, it influenced the thinking of many scientists working in the field of psychology, too. One of the basic notions of the theory of evolution was that members of the same species are not alike, that is, individual differences exist among members of a species. Sir Francis Galton, a half-cousin of Charles Darwin, was quick to pick up the idea of individual difference and gathered data on the individual differences of physical and psychological characteristics. He wanted to measure the basic individual differences among human beings and for this he singled out 'human ability' as a possible dimension of study and, in doing so, laid the foundation stone of the branch of psychology which is today known as psychometrics. In fact, Galton was obsessed with the idea of measurement and had firm belief that virtually anything was measurable. His attempt to measure intelligence by means of reaction time (RT) and sensory discrimination tasks were well-known. His most influential works were *Hereditary Genius* (1869) and *Inquiries into Human Faculty and its Development* (1883). The former book emphasized the genetic factors which were held to be important for eminence and the latter was a series of essays that emphasized individual differences in mental faculties. With the help of time-consuming psychophysical methods practiced by Wundt and others, he tried to assess simple sensory-motor responses. Thus, he continued the tradition of brass instrument mental testing but with one difference: his procedure was modified in such a way that it was amenable to data collection from hundreds, if not thousands, of the testees. Due to his efforts in popularizing practicable measures of individual differences, Galton is regarded by most historians as *father of mental testing* (Goodenough 1949).

To continue his study of individual differences, Galton set up a psychometric laboratory in London at the International Health Exhibition in 1884. Here the tests and measures being administered involved both the physical and behavioural measures. However, in the long run, Galton's attempt to assess intellect with the measures of reaction time and sensory discrimination proved to be a fruitless effort.

Later, Karl Pearson developed several sophisticated techniques for statistically analyzing the data on individual differences of human ability, which provided a good stream for psychological measurement. The product-moment method of computing the correlation coefficient was one such technique. J M Cattell, who had studied with both Wundt and Galton, continued testing and measurement programme in a scientific way. He invented the term 'mental test' in his most important paper *Mental Tests and Measurement.* He started his own research laboratory at Columbia University, at that time the largest university in America. There, he developed a series of tests that were mainly extensions and additions to Galton's theory. In fact, the physiological and sensory bias of his various test battery reflects strongly of its Galtonian heritage. Thus, Cattell tried his best to import the brass instruments approach to America. Among the most influential students of Cattell were E L Thorndike, R S Woodworth and E K Strong who, by devising various tests, kept the spirit of psychological testing and measurement continue. But among Cattell's student, it was probably Clark Wissler who showed some surprising results in 1901 and therefore, had the greatest influence on the early history of testing and measurement. In fact, his aim was to demonstrate that the test scores could predict academic performance. His analysis showed virtually no tendency for mental test scores to correlate with academic performance, which thereby redirected the mental testing movement away from brass instruments. Also damaging to the brass instrument testing movement was very low correlations between the mental tests themselves. For example, RT and colour naming correlated –0.15 and colour naming and hand movement correlated only 0.19. Thus with the publication of such discouraging results, the experimental psychologists finally dropped the idea of using RT and sensory discrimination as measures of intelligence. One point of view is also that this turning away from the brass instrument approach was important as it paved the way for immediate acceptance of Alfred Binet's useful measure of human ability.

Apart from the above main streams of psychological measurement, clinical studies growing out of medicine, psychiatry and social welfare institutions in western countries also provided immense impetus for the growth of measurement. Clinical interest in the feeble-minded, insane and misfit was first evoked in about 1900 in France. Psychologists working there were deeply interested in developing tests and instruments which could measure maladjustments and identify the possible reason for insanity. Alfred Binet in 1905 gathered data based upon the intelligence of school-going children in this context, which was later on manifested in terms of the development of different intelligence scales.

The beginning of the twentieth century was marked by a tremendous growth of psychological and educational tests, and the development of measurement tools and techniques started on a war footing. The first psychological test to appear was the Binet–Simon Intelligence Scale. It proved to be one of the most promising verbal scales for measuring the intelligence of children. It was a 30-item measure of intelligence. Curiously, there was no method of scoring the test (Gregory 2005). It simply aimed at identifying those school-going children who could not benefit from regular instructions in class. It has been revised several times. Its 1912 revision was important because in this revision Stern introduced the concept of IQ.

Later on, its several American revisions appeared. Besides this, several other achievement tests including the famous Thorndike Handwriting Scale appeared. In 1918, Woodworth's Personal Data Sheet appeared, which was the first personality inventory. During the two World Wars, the two well-known tests—the Army Alpha test and the Army Beta test—were developed. Several group intelligence tests and achievement tests also made their appearances. The outcome of all these enthusiastic developments in the sphere of psychological tests was a rapid development of psychological and educational measurements. During the last 30 years or so, a different trend has become evident. Much emphasis is now being put upon the methodology of development and use of measurement in almost all branches of psychology. Newer methods of development of psychological and educational tests have become the main centre of interest for psychologists. They are frequently found criticising the old methods and placing arguments for the adaptation of a new method. This methodology is known as the *psychometric theory* which mainly includes the basic principles of test construction.

For a still better understanding, the history of mental testing in 20th century can be sketched and explained by dividing it into six periods (Thorndike and Thorndikechrist 2015): *early period, boom period, first period of criticism, battery period, second period of criticism and the period of accountability.* A brief discussion follows.

(i) *The Early Period:* This period extends from 1901 to 1915 and was, in fact, a period of tentative exploration and theory development. The Binet–Simon scale was published in 1905 and was revised twice by Binet. The test was so popular that it was brought to the USA by several pioneers in the field of mental testing and measurement. Of the various pioneers, Lewis Terman of Stanford University was most influential. In 1916, he published the first version of the test and ularly called as Stanford–Binet intelligence scale. One of the colleagues of Terman, Arthur Otis, began work in testing mental ability of children and adults in group. S D Portens also did a commendable work in Australia by developing a maze test of intelligence to be used mainly with people having hearing or language handicaps.

Another influential person in this field was a British psychologist, Charles Spearman, who developed two theories relating to the measurements of human abilities. One theory was the statistical theory which tried to describe and account for the inconsistency in the measurements of human behaviour. This theory has gradually developed into the concept of reliability. Spearman's another theory stated that there is a single dimension of ability underlying most human performance and the consistency of people's performance on different ability measures was simply the result of this general intelligence (or popularly called as g-factor of intelligence).

(ii) *The Boom Period:* This period started with American involvement in World War I. In fact, this involvement created an urgent need to improve the functioning of the army and psychologists were considered the most fit persons for enhancing this improvement programme. Such an event and the recognition of the role of psychologists created a 16-year *boom period* (that is up to almost 1930) which witnessed many types of advances and innovations in the field of mental testing and measurement. A group of psychologists led by Robert Yerkes expanded Otis's work for developing and implementing large-scale group testing of ability. Two popular tests, the Army Alpha test (a verbal test) and the Army Beta test (a performance test), were developed. These tests were extensively used for selecting officer trainees as well as for removing intellectually poor military personnel from military service. For identifying emotionally unstable and therefore, unfit for military service, the first objective measure of personality called Woodworth Personal Data Sheet, was developed in 1918.

After the war, psychologists showed interest in studying a variety of behaviours such as interest and attitudes. E K Strong and some of his colleagues developed Strong Vocational Interest Blank (SVIB) shortly after World War I for assesing vocational interests to help college students choose careers consistent with their interests. Similarly, in 1929, L L Thurstone proposed methods of scaling for measuring attitudes and values. However, the period following World War I witnessed a low point for the history of mental testing movement because some critics like Walter Lippmann criticised both these tests as well as the conclusion drawn from the test scores.

(iii) *The First Period of Criticism:* This period roughly covered the period of 1930s and has been described as a period of *Criticism and Consolidation* . Many new tests were published and some were revised during this period. The most popular personality test, the Minnesota Multiphasic Personality Inventory (MMPI), was developed and Wechsler-Bellevue Intelligence Scale also came up during this period. Kuder Preference Record was also developed. Some advances were also made in mathematical theory underlying test. For example, L L Thurstone refined the technique of factor analysis and the theory underlying it.

However, some voice was raised against the increasing use of test scores for making different types of decisions. One young psychologist, Oscar Buros, clearly observed that many psychological and educational tests had no scientific and objective evidence for supporting their uses. In 1935, he started the *Mental Measurement Yearbook* (MMY), which served as a place of publication for critical reviews of these tests and their functioning. In fact, he intended to get reviews of tests from the experts so that these tests can be further refined, their uses may be more objectified and supported.

(iv) *The Battery Period:* During the 1940s and 1950s, again, widespread use of psychological and educational tests was started. Based on the theory of Thurstone and others, there were several dimensions of abilities, some test batteries were developed. The success of these test batteries encouraged the movement of testing further and experts were showing much interest in developing test batteries and related statistical technique called factor analysis.

Taxonomies of ability such as those provided by Bloom and Guilford were found much helpful in describing the range of various types of mental functioning.

During the 1950s, psychological and educational testing became much like a business concern. Many test batteries were being used for assessing academic achievement of students in the class. Business organisations, industry and the civil service system increasingly made the use of personality, ability and aptitude test for assessing various mental functioning. In this direction, the use of the Differential Aptitude Test (DAT) and the General Aptitude Test Battery (GATB) were most common. Not only this, in mental hospitals, patients were routinely assessed through various personality tests and ability tests. In 1954, the American Psychological Association also released a set of guidelines for effective use of various psychological and educational tests.

(v) *The Second Period of Criticism:* The period started in the 1960s when there was underlying revolt against male dominance in American society and tests were considered as a biased tool of male oppression. At that time, on most ability tests, the scores of females were lower than the scores of males. Likewise, personality tests were also thought as unnecessarily favouring the male personality. At that time, the most serious concern was the possible use of ability and personality tests in making discrimination against females or members of minority groups in employment and education. As a consequence, tests were being subjected to various criticisms. Some legal proceedings in the courts of law were also started in the USA.

(vi) *The Period of Accountability:* Despite the fact that there was a public criticism of testing and measurement tools, the government in the USA relied more and more on the use of standardized psychological and educational test in assessing school achievements of students. In 2002, the NCLB (No Child Left Behind) bill was passed by the US government that ensured that all public school students must achieve high standards of performance in areas like science, mathematics, arts, etc. Low-performing schools were subjected to a variety of interventions ranging from technical assistance to various types of restructuring programmes. Schools must be held accountable for the performance of students and they can assess their performance by using standardized tests.

LEVELS OF MEASUREMENT (OR MEASUREMENT SCALES)

Before we go on to discuss the different levels of measurement scales, we must know something about the different postulates of measurement. A postulate is defined as a sort of assumption which stipulates the relationship between groups, objects or events being measured. The postulates will help us to understand the different relationships among the objects measured by the different levels of measurement scales. According to Guilford (1954, 11), there are nine basic postulates of measurement. These are briefly summarized under the following three general headings.

1. Postulates relating to equalities or identities
2. Postulates relating to rank order
3. Postulates relating to additivity

Ordinarily, there are three variations in postulates relating to equalities or identities, two variations in postulates relating to rank order and four variations in postulates relating to additivity. Thus a total of nine postulates of measurement exist.

The three variations in postulates relating to identities or equalities are as follows:

1. Either $a = b$ or $a \neq b$. This means numbers are either equal ($a = b$) or they are not equal ($a \neq b$) but not both. This postulate is more essential for classification.
2. If $a = b$ then $b = a$. This means that the relation of equality among numbers is symmetrical, and therefore, we can interchange a for b or b for a.
3. If $a = b$, $b = c$ then $a = c$. This means that the objects which are equal to the same object are also equal to one another.

The two postulates relating to rank are:

4. If $a > b$ then $b \ngtr a$. This means that the relation between a and b is asymmetrical, that is, we cannot reverse the relationship $a > b$ or $a < b$ interchanging a and b.
5. If $a > b, b > c$ then $a > c$. This is a transitivity postulate and indicates a definite order in the ranking. It is an important postulate upon which most measurements in psychology, sociology and education are dependent.

The four postulates relating to additivity are as follows:

6. If $a = p$ and $b > 0$ then $a + b > p$. This postulate indicates the summation process and also demonstrates that when a zero is added, it produces no variability in the result.
7. $a + b = b + a$. This postulate denotes that in the process of addition, the order of the numbers is not important because if $a = 4, b = 2$ then $4 + 2$ is equal to 6, and $2 + 4$ is also equal to 6.
8. If $a = p$ and $b = q$ then $a + b = p + q$. This postulate indicates that in the process of addition, identical numbers or objects may be substituted for one another without making any change in the result.
9. $(a + b) + c = a + (b + c)$. This postulate indicates that in the process of addition, the order of combinations of objects or numbers makes no difference.

As we know, measurement is the process of assigning numerals to the attributes of objects according to some rules. Following Stevens (1951), there are at least four different ways of assigning numerals to the attributes of the objects. Based upon these four ways, there are four *different levels of measurement* (or scales). The first five postulates are enough to account for the following four levels of measurement.

Nominal or Classificatory Scale of Measurement

Nominal measurement (or scale) is the *lowest* level of measurement. In nominal measurement numbers are used to name, identify or classify persons, objects, groups, etc. Nominal scales are really not scales and their only purpose is to name objects. For example, a sample of persons being studied may be classified as (a) Hindu, (b) Muslim and (c) Sikh or, the same sample may be classified on the basis of sex, rural–urban variable, political party affiliation, etc. All these classifications would be examples of nominal measurement. Classification of persons into clinical groups such as schizophrenia, manic-depressive psychosis, phobia, etc., also constitutes nominal measurement. In nominal measurement, members of any two groups are never equivalent but all members of any one group are always equivalent. And this equivalence relationship is reflexive, transitive and symmetrical. Nominal data are counted data. In case of nominal measurement admissible statistical operations are counting or frequency, percentage, proportion, mode, and coefficient of contingency. Addition, subtraction, multiplication and division are not possible because the identifying numerals themselves cannot be legitimately added, subtracted, multiplied or divided. The drawback of nominal measurement is that it is most elementary and simple. It is the least precise method of quantification. Because of these characteristics, some experts are of the view that nominal measurement is not a measurement at all.

Ordinal Scale of Measurement

This is the *second* level of measurement in which there is the property of magnitude but not of equal intervals or an absolute zero. In ordinal measurement, numbers denote the rank order of the objects or the individuals. Here numbers are arranged from highest to lowest or from lowest to highest. Ordinal measures reflect which person or object is larger or smaller, heavier or lighter, brighter or duller, harder or softer, etc., than others. Persons may be grouped according to physical or psychological traits to convey a relationship like 'greater than' or 'lesser than'. Socio-economic status is a good example of ordinal measurement because every member of the upper class is higher in social prestige than every member of the middle and lower class. Likewise, every member of the middle class is higher in social prestige than every member of the lower class. Students may be ranked 1st, 2nd and 3rd in terms of their academic achievement to constitute the example of ordinal measurement. Thus, in ordinal measurement, besides the relationship of equivalence, a relationship of 'greater than' or 'lesser than' exists because all members of any particular subclass are equivalent to each other and at the same time greater or lesser than the members of other subclasses. The relationship of greater than is usually irreflexive, transitive and asymmetrical. The persmissible statistical operations in ordinal measurement are median, percentiles and rank correlation coefficients, plus all those which are permissible for nominal measurement.

The drawback of ordinal measurement is that ordinal measures are not absolute quantities, nor do they convey that the distance between the different rank values is equal. This is because ordinal measurements are not equal-interval measurements, nor do they incorporate absolute zero point. Let us take an example. Suppose an honesty test is given to a sample of 30 students. A and B are given the rank of 3 and 7 respectively, whereas D and Z are given the rank of 11 and 15 respectively. In this example, B ranks four points above A, and similarly Z ranks 4 points above D. But from this we cannot say that the difference between A and B, and that between D and Z is equal because in ordinal measurement, rank values which may be equally spaced do not convey that the underlying properties will also be equal. Likewise, intelligence scores such as 50, 40, 20 of X, Y and Z respectively may be ranked as 1, 2 and 3 respectively. The difference between 50 and 40 is 10 and the difference between 50 and 20 is 30. But both these differences are equally spaced in terms of ranks. Thus in ordinal measurement from equally-spaced rank values, it is difficult to infer that the underlying trait will also be equally spaced. The second demerit of ordinal measurement is that there is no way to ascertain whether a person has any of the characteristics being measured.

Interval or Equal-interval Scale of Measurement

This is the *third* level of measurement and includes all the characteristics of the nominal and ordinal scales of measurement. The salient feature of interval measurement is that numerically equal distances on the scale indicate equal distances in the properties of the objects being measured. In other words, here the unit of measurement is constant and equal. This is the reason why interval measurement is also referred to as *equal-interval* measurement. Since the numbers are after equal intervals, they can legitimately be added and subtracted from each other. For example, suppose four objects A, B, C and D have been measured and given the score of 20, 16, 8, 4 respectively on an interval measurement. Here the difference between $A - C = 20 - 8 = 12$ is equal to $B - D = 16 - 4 = 12$. Interpreting the same in still another way, it can be said that the interval between A and $B = 20 - 16 = 4$ and the interval between C and $D = 8 - 4 = 4$. These intervals can be added together: $(20 - 16) + (8 - 4) = 4 + 4 = 8$. Thus on an interval measurement, the intervals or distances (not the quantities or amounts) can be added. Therefore, in an interval scale the difference (or interval) between the numbers on the scale reflects difference in magnitude. However, the ratios of magnitudes are meaningless.

Although it is true that the intervals can be added, it does not mean that the process of additivity can be carried out in the absolute sense. The process of additivity of intervals or distances on an interval measurement has only a limited value because in such a measurement, zero point is not true but rather arbitrary. Zero point, here, does not tell the real absence of the property being measured. It is selected only for some convenience in the measurement. Suppose that the two values on an interval measurement are 5 and 8. Obviously, these two values mean that 5 and 8 are 5 and 8 points above the arbitrary zero point and, adding these two, we get 8 + 5 = 13, which means that we have a total quantity, that is, 13 points above the arbitrary zero point. Now, suppose for the time being that this arbitrary zero is itself 3 points above the real zero point, then in that case the value 5 should be 5 + 3 = 8 and the value 8 should be 8 + 3 = 11, if the same meaning is to be conveyed. Adding 8 + 11, we get 19 and not 13. Thus the process of additivity on an interval measurement cannot be carried out in the absolute sense of the term. Fahrenheit and Celsius thermometers, time as shown in our calendar, scores on intelligence tests and aptitude tests are good examples of interval measurements. If a person secures a score of zero on any numerical aptitude test, it does not mean that he has no knowledge of numerical operations. Likewise, if a person secures a score of 80 on the same numerical aptitude test, it does not mean that he has a numerical aptitude which is twice greater than the person who has secured 40 on the same numerical aptitude test. All this happens because the zero point on the numerical aptitude test is not true, rather it is arbitrary. Psychological tests and inventories are the interval scales and therefore, have this limitation. The common statistics used in such measurement are arithmetic mean, standard deviation, Pearson *r*, and the other statistics based upon them. Statistics like the *t*-test and *F*-test, which are widely used tests of significance, can also be legitimately applied. The only statistics which cannot be applied in interval measurement is the coefficent of variation. The reason is that the coefficient of variation is a sort of ratio of standard deviation to the arithmetic mean. Standard deviation is a fixed deviation on the measurement scale because it is not affected by any shift in the zero point. But the mean is likely to vary whenever there occurs a shift in the zero point. When the mean is affected, the coefficient of variation will also be affected. As such, it is advisable *not* to calculate the coefficient of variation from interval measurements.

Ratio Scale of Measurement

It is the *highest* level of measurement and has all the properties of nominal, ordinal and interval scales plus an absolute or true zero point. The salient feature of the ratio scale is that the ratio of any two numbers is independent of the unit of measurement and therefore, it can meaningfully be equated. For example, the ratio 16:28 is equal to 4:7. In ratio measurement, all the nine postulates of measurement can be applied. Also, all statistical operations including the coefficient of variation can be utilized.

The common examples of ratio scale are the measures of weight, width, length, loudness, and so on. It is obvious, therefore, that ratio scales are common among physical sciences rather than among social sciences. To be more clear, a student must know the distinction between the interval scale and the ratio scale. The fundamental difference is that in the former, the zero point is arbitrary but in the latter, the zero point is true. Temperature measured in terms of Fahrenheit and Celsius is an example of interval scale and length measured in terms of feet and inches is an example of the ratio scale. When we measure the length of two sticks and say that one stick is 3′ long and another 6′ long, we have a clear idea that the second stick is twice the length of the first stick. This is because 6′ means actually 6′ from 0′, and 3′ means actually 3′ from 0′. But suppose the maximum temperature of weather today is 40°F and it was 20°F last year on the same date. Can we say, now, that today is twice as warm as it was on the same date last year? Obviously, the answer is no because 0°F does not reflect the complete absence of temperature (true zero point) in Fahrenheit measurement.

In psychology, sociology and education we frequently encounter interval measurement because most of the data obtained from measurement in these sciences are such that we assume an equal unit of measurement and arbitrary zero point. The basic properties of all these four scales of measurement has been summarized in Table 1.1.

Table 1.1 Basic properties of scales of measurement

Types of scales	Property		
	Magnitude	Equal intervals	Absolute zero
Nominal	No	No	No
Ordinal	Yes	No	No
Interval	Yes	Yes	No
Ratio	Yes	Yes	Yes

PROPERTIES OF SCALES OF MEASUREMENT

There are three important properties that make scale of measurement different from one another—*magnitude, equal intervals* and *absolute zero*. These are discussed below.

1. *Magnitude:* In fact, magnitude is defined as the property of "moreness". Any scale is said to have the property of magnitude if it can be said that a particular instance of the attribute represents more, less or equal amounts of the given quantity than does another instance. (Mc Call 1994). For example, on a scale of weight, if it can be said that Mohan is heavier than Sohan, then the scale can be said to demonstrate the property of magnitude. But suppose the experimenter labels his samples as Group I, Group II and Group III, and makes no reference of moreness, the scale will not be said to have the property of moreness.

2. *Equal intervals:* A scale has the property of equal intervals if the difference between any two points at any place on the scale has the same meaning as the difference between two other points that differ by the same number of scale units. For example, the difference between 4 kilograms and 6 kilograms on a weight-measuring scale represents the same quantity as the difference between 16 kilograms and 14 kilograms, i.e., exactly 2 kilograms.

The above simplicity of meaning, however, does not imply to a psychological test, which rarely has the property of equal intervals. For example, the difference between IQs of 50 and 60 does not mean the same thing as the difference between IQs of 110 and 120. Although each of these differences is of 10 points, (that is, 60 – 50 = 10 and 120 – 110 = 10), the 10 points at the first level do not mean the same thing as 10 points at the second level. In fact, when the scale really has the property of *equal intervals*, the relationship between the measured units and some outcome can be described by a straight-line equation in the form of $y = a + bx$. This equation obviously shows that an increase in equal units on a given scale reflects equal increases in the meaningful correlates of the units (Kaplan & Saccuzzo 2001)

3. *Absolute zero:* An absolute zero is said to exist when nothing of the property being measured exists. For example, if a doctor is measuring the heart rate of a patient and finds that the patient has a heart rate of zero and has died, the doctor would readily conclude that there is no heart rate at all. But for many psychological measurements it is very difficult, if not impossible, to define absolute zero. For example, if a psychologist wants to assess shyness on a scale from 0 through 5, it is very hard to define what it means for a person to have absolutely no shyness.

These are the basic attributes of scales of measurement over which all the four scales of measurements, as pointed out by Stevens, differ from each other.

FUNCTIONS OF MEASUREMENT

In psychology and education, measurement has got varied functions. It may help a class teacher to distinguish between the high achievers and the low achievers; it may help a learning psychologist to diagnose the current status of the reading ability of a particular child; or it may help an industrial psychologist to place a worker in the right job. The primary functions of measurement are as follows.

In Selection

Selection of personnel in industry or other institutions may be carried out with the help of psychologists. The function of measurement tools in selections is to predict the ability of the individual. Sometimes it is found that the personnel selected fail to prove their mettle and the management is forced to take some drastic action against them such as dismissal, suspension, etc. Here again, measurement plays an important role by providing data on the basis of which suitable disciplinary action can be recommended against such personnel.

In Classification

Measurement also helps in various types of classification, which sometimes becomes very necessary for a programme to be carried out effectively. For instance, there may be some problem children in a class and their class teacher may wish to separate them in order that the other children in the class do not suffer. Here, measurement will help the teacher in classifying the children into the following categories: retarded, average and gifted. Likewise, a psychiatrist or a clinical psychologist may wish to classify his patients into different categories of mental illness and here again, he is helped by the tools of measurement. In industry also, measurement helps personnel to be classified according to the indices of job satisfaction, accident proneness, absenteeism, etc.

In Comparison

The pioneering work of Galton and Darwin has revealed that no two individuals are alike. In other words, there always exist individual differences in traits, mental processes, habits, tendencies, educational achievements, abilities, etc., between any two individuals. Whenever two persons are to be compared on any of the above factors or so, measurement comes into use. With the help of appropriate measurement, it is possible to conclude that A is better than B in spelling tests or that students of class X of an urban school are better in their academic achievements than their counterparts in a rural school.

In Guidance and Counselling

Measurement also assists psychologists in guiding and counselling. Counselling is a sort of specialized guidance programme and refers to the advice given to an individual so that he can arrive at a workable solution to various adjustment problems in life. Measurement can help an individual to know his strengths and weaknesses; it may provide an insight and understanding into the relationship between the counsellor and the patient; it can help to make accurate predictions regarding problems of adjustment likely to come up in the future and also diagnose mental disabilities, aberrations, deficiencies, etc.

In Research

Measurement helps in research activities, too. In fact, measurement is the fundamental basis of all psychological and educational researches. Research is the investigation undertaken to discover new facts about a problem. In psychological and educational researches, usually, the effect of one variable or a set of variables is studied while the effects of all other variables are controlled. In doing so a variety of measurements is resorted to, to confirm that the design, statistical calculation, results, etc., are in accordance with the principles of measurement or else research becomes meaningless. Thus, measurement may be regarded as the heart of any psychological or educational research.

In Improving Classroom Instruction

A teacher teaches all the students in the classroom in a similar manner. However, the results are different for the various students. Obviously, for some students the instruction is at par with their mental ability but for others it may be either below or above their mental ability. In both the cases, there is a need for modification in the instruction, both quantitatively as well as qualitatively. To achieve this, measurement plays a vital role. By measuring the outcome of the instruction, it becomes clear how many students are being benefited by the instruction and how many are not being benefited by it or are least benefited by it. After such an analysis some essential suggestions for modification in the instruction can be made.

DISTINCTION BETWEEN PSYCHOLOGICAL MEASUREMENT AND PHYSICAL MEASUREMENT

Measurement has two broad dimensions—psychological or qualitative measurement and physical or quantitative measurement. Psychological measurement comprises the measurement of mental processes, traits, habits, tendencies, and the likes of an individual whereas physical measurement comprises the measurement of objects, things, etc., which are often physically present in the world. Usually, physical measurement is concerned with the measurement of height, weight, length, size, volume, etc. Given below is a clear-cut distinction between psychological and physical measurement.

1. In physical measurement, the unit of measurement is fixed and constant throughout the measurement whereas in psychological measurement the unit of measurement is not fixed and varies during the process of measurement. For example, a kilogram or an inch has the same meaning at whatever place the measurement is being taken and it conveys the same physical significance or meaning throughout the measurement. But suppose one is measuring intelligence. There is no fixed unit of measurement in this case because some may measure intelligence on the basis of verbal questions or items answered in a specified time; others may prefer to measure intelligence on the basis of some manipulative tasks done in a specified period; still another group may prefer to measure intelligence on the basis of both time and error in the completion of a task, and so on. Moreover, these units tend to vary themselves during the process of measurement because there is no standard method of presenting a uniform set of difficulties to all examinees. For example, a particular item to measure intelligence may seem very easy to one examinee but may seem very difficult and challenging to another.

2. In physical measurement, there is a true zero point whereas in psychological measurement there is an arbitrary zero point. By a true zero point is meant a point which actually represents the underlying absence of the trait being measured whereas by an arbitrary zero point, we mean a point which does not represent the underlying absence of the trait being measured. For example, when a person gets a score of zero in a numerical ability test, it does not mean that he has no knowledge of numerical operations at all. But an object having zero length will be said to have no length at all.

3. Physical measurement is more accurate and predictable than psychological measurement. This is because in physical measurement, there is a true zero point. For example, a stick of 20" length means the stick is definitely 20" above the zero inch, and similarly, a stick of 60" length means that it is thrice as long as the first stick. But a person scoring point 15 in an intelligence test cannot be said to have scored 15 points above the zero point because here zero point is itself not known. Likewise, another person securing score 30 cannot be said to be twice as intelligent as the first person. Therefore, no prediction can be made with definite accuracy.

4. Physical measurement is direct whereas psychological measurement is indirect. Whe we want to measure the length of a cloth, we place it before us and directly measure its length inches or feet. However, it is not possible to measure the 'extroversion' trait of personality intelligence in a like manner as it cannot be placed physically before us. Extroversion can only measured indirectly through some responses given by the person concerned and for measuri intelligence, we will have to depend upon some responses—verbal or manipulative.

5. In physical measurement the entire quantity can be measured, whereas in psychological measurement the entire quantity cannot be measured but only a sample representing that quantity or trait is measured. Say, for instance, one is to measure the lengths and weights of all tables and chairs in one's home. The entire lengths and weights of all the tables and chairs can be measured and expressed in terms of inches and kilograms respectively. But suppose one is measuring the mechanical aptitude of class X students of Bihar. Then ordinarily, it is not possible to measure the mechanical aptitude (through an appropriate test) of every student of class X belonging to the state of Bihar. Naturally, one would randomly draw a sample of students who are taken to be representative of class X and measure their mechanical aptitude.

STEPS IN MEASUREMENT PROCESS

Psychologists and educators have shown keen interest in the measurement of human abilities, interests, attitudes and personality traits. Whatever is being measured in any field, the measurement involves three common steps:

(i) Identifying and defining the quality or attribute

(ii) Determining operations for isolating and displaying attribute

(iii) Quantifying the attribute.

A discussion follows.

(i) *Identifying and Defining the Quality or Attribute* The first step in measurement is to identify the attributes that are relevant and important to measure. For example, if a company is selecting employees to become truck drivers, the ability of verbal comprehension and ability to solve quantitative problems will not be relevant attributes. However, other abilities or functions such as eye-hand coordination, depth perception, low accident records will be the relevant attributes.

After selecting the relevant attribute, the next step will be to define the attribute. Psychologists and educators face great difficulty in defining abstract and difficult-to-observe property of the people. Such property is technically called as *construct*. For example, intelligence is a construct. Frequently asked questions here are: What kind of behaviour shall be characterized as intelligence? Will it include skill in social interaction? Will it relate to ability to deal with only concrete objects? Will it refer to ability to adjust with novel situation? Will it refer to the speed and fluency of a response?, etc. A psychologist or educator must arrive at a clear, precise and generally accepted definition of the attribute considered as relevant one.

(ii) *Determining Operations for Isolating and Displaying Attribute* The next step in developing a measurement procedure is how to find a set of operations that will isolate the attribute of interest and correctly display it. As we know, when an attribute is defined by how it is measured, it is said to have an operational definition. In so far as the physical measurement is concerned, this step is not much difficult and the person easily determines those operations through which the attribute can be measured. For example, if one is to measure the length of the table, the ruler and the tape measure are uniformly accepted instruments and laying one of them along the table is the correct procedure for assessing the length of the table. But in case of psychological and educational measurement, this process is a bit complicated, because there have been several types of invention of instruments and procedures for eliciting behaviours that serve as indicators of the relevant attributes of the people. For example, how do we operationally define intelligence? What behaviours are indicative of the attribute—intelligence? The fact that there is no single, universally accepted test of intelligence and that different tests include different tasks for defining intelligence, becomes evidence for the fact that we don't have complete consensus on what appropriate procedures are for eliciting intelligent behaviour. This lack of consensus is the general characteristic of psychological and educational measurement and is the outcome of ambiguity in the definitions of the attribute and the varieties of instruments devised for eliciting the relevant behaviours of the attribute.

(iii) *Quantifying the Attribute* Once the set of operations for identifying an attribute has been accepted, the next step is to express their outcomes in quantitative terms. This is usually done by assigning some numerical values according to a set of rules. This set of rules is called scale, which allows to answer questions such as how many? or how much? In case of length of a table, the question basically becomes, how many inches? how many feet? In these physical measurements, any two inches are equal and this is established by direct comparison. However, in case of psychological and educational measurement of attributes, there are no such units whose equality can be demonstrated by direct comparison in the way that equality of inches is shown. Psychologists or educators have to depend on a somewhat arbitrary definition to provide units and quantification. Generally, a researcher considers one task successfully completed such as an arithemetic problem solved, a word defined or an attitude statement endorsed equal to any other task and subsequently, use the total number of successes or endorsements for a person as the correct representation of value representing the person on that attribute. Such count of tasks successfully completed provides a manageable definition of the amount but has no adequate evidence of equivalence of different tasks in the series.

As said above, in quantifying the attribute, numerical values are used. Using numerical values in measurement has several advantages of which two are important ones.

(a) Use of numerical values in measurement facilitates communication and makes it meaningful and precise. For example, when we say that Mohan is 6 feet tall or the weight of Ram is 60 kg, a clear meaning regarding height and weight is communicated.

(b) Another advantage of quantification is that it allows useful mathematical operations in summarizing the quantitative information. For example, by adding, subtracting, multiplying and dividing several information, we find the average to convey the summary of the performance.

These are the major steps in the measurement process that a psychologist or educator has to go through.

PROBLEMS RELATED TO THE MEASUREMENT PROCESS

We have seen that there are three steps in the measurement process. In psychological and educational measurement, several problems relating to each of these three steps are encountered.

(i) The researcher encounters problems in selecting the attributes of interest and in defining them clearly and unequivocally. In defining concepts like intelligence, anxiety, adjustments, cooperativeness, and so on, we expect diversity in definition. For example, in defining intelligence, to what extent a definition should include each of the following abilities:

- ability to memorize
- ability to profit by past experience
- ability to carry out abstract reasoning
- ability to make adjustment in various situations
- ability to generalize from an event

(ii) The second problem relates to devising procedures for eliciting the relevant attributes. For some attributes, the psychologists or educators have succeeded in identifying suitable measurement operations. For example, in the domain of abilities, they have been able to develop various standardized tests through which measurement is done easily and objectively. But, for some other attributes such as for a client's anxiety, fear, the initiativeness and cooperativeness of the employee no such satisfactory standardized measures are available. It is hoped that in the future psychologists or educators, through their active research, will be able to find out some standard operations for assessing these attributes also.

(iii) The third problem relates to the equality of units. In psychological and educational measurements, units are set equal by definition and in reality, their equality cannot be established in fundamental sense. In a foot, there are twelve inches and all inches or units of measurement are equal. But in intelligence tests, items of analogy such as 'Hot is to cold as wet is to...?, are not equal to items of arithematical series such as 2, 8, 13, 17,...? Such inequality creates problem in the measurement process. Perhaps due to such inequality, addition, subtraction and comparison of scores remain a suspect issue.

Despite all these problems in the psychological and educational measurement process, it has been found that the information provided by tests is more accurate than information available from other sources.

GENERAL PROBLEMS OF MEASUREMENT

Science is a kind of social institution. Its scientific value lies in meaningfully communicating the observations of events or individuals to other persons. It is towards this end that measurement, which is a sort of quantitative description of events or individuals, makes a significant contribution. As a matter of fact, the progress of any discipline is today judged by the extent to which it has been able to make a quantitative description (i.e., measurement) of its subject matter. The reason why we attach such importance to measurement is that it allows extremely accurate and objective quantitative description of events. In fact, accuracy and objectivity are its two principal advantages.

Despite the fact that measurement is the heart of the social sciences, it has certain basic problems which social scientists need to look into. Measurement would not be of much importance till these difficulties have been removed. Some of these problems are enumerated below.

Indirectness of Measurement

Most psychological and educational measurements are indirect. This is because most psychological and educational variables cannot be observed and studied directly. For example, suppose a teacher wants to measure the intelligence of students of a particular class. As intelligence cannot be directly seen, touched, or experienced, the teacher has to depend upon measures which include a sample of behaviour representative of an intelligent act. Such a sample, however, may itself suffer from a number of limitations. For example, these measures may not be reliable, valid, and practical or may not be objective and true representative of the actual behaviour being measured. In such a situation, measurement of any trait or variable itself becomes a source of perpetual difficulty.

Incompleteness of Measurement

Psychological and educational measures are generally incomplete, and, therefore, the measurement of any psychological or educational variable is also incomplete. For example, when an investigator is assessing the attitude towards co-education, he is required to construct a scale in which a number of samples of behaviour expressing such an attitude need to be incorporated. This number has no limit. Any attempt, therefore, to measure such an attitude would be partial and incomplete. In such a situation, measurement will be dubious and tend to create a misleading index of attitude.

Relativity of Measurement

Psychological and educational measurements are relative. This is also true of sociological measurement. The concept of relativity in measurement can be explained through an example. Suppose Mohan, a student of class X, was given two tests—a test of arithmetical knowledge and a test of the English language. Let us further suppose that he correctly answered 60% of the items in the English language test but could not answer even a single item in the arithmetic test. On the basis of such a measurement, can we say that Mohan's performance in the English language test was good? The answer obviously cannot be given with all certainty. 60% of the items answered by Mohan may be the percentage for even those students of the class who were much below the average. On the other hand, the test of English language may have been a very difficult test and only Mohan and nobody else might have answered 60% of the items correctly. Thus, the measurement is not absolute but relative and we cannot draw any inference from the measurement of Mohan's performance unless it is compared with the reference group, that is, with other members of the class. Likewise, can we say, on the basis of his performance on the arithmetical knowledge test, that Mohan has no knowledge of arithmetical operation? We cannot say so because the zero obtained in arithmetical knowledge test does not reflect the absence of arithmetical knowledge. Nothing can be definitely said until a comparison with other members of the class is done. All these measurements are, therefore, relative and must be carefully dealt with if measurement is to be meaningful and objective.

Errors in Measurement

Measurement in the physical sciences as well as in the behavioural sciences is most of the time not pure. It contains some uncontrolled factors, which produce gross errors. Suppose a weighing machine determines a woman's weight to be 50 kg. This weight might not be her pure weight. There may be some minor mechanical troubles in the machine itself so that her weight is inflated; it may be that she has just taken her meal; it may be that she is pregnant; there may be other factors present in the physical environment. All these sources of error might inflate or reduce her actual weight. Similar sources of error run into psychological, educational and sociological measurement. When we are measuring the intelligence of a child with the help of an intelligence test, there can be several such factors which tend either to decrease or increase his actual score. For example, the child might be nervous; he might have been distracted by the sound of an aeroplane; he might not have understood the meaning of the items clearly, and so on. All these sources of errors in measurement create problems which adversely affect the scientific value of measurement.

SOURCES OF ERRORS IN MEASUREMENT

In any ideal research study the measurement should be precise, scientific and unambiguous. But this goal is not always reached because of the various sources of errors often encountered. The following are the important and possible sources of errors in measurement:

(i) *Respondent*: One important source of errors in measurement is the respondent himself. Sometimes the respondent is found reluctant to express his true feelings or it may be that sometimes due to lack of knowledge, the person may not express himself clearly. In either case, the measurement loses its accuracy.

(ii) *Measurer*: Sometimes the behaviour, style and looks of the person, who is measuring the phenomena, distorts the process of measurement. His behaviour, style and looks may encourage or discourage certain types of replies from the respondent that affects the accuracy of the measurement. At the data-analysis step, due to incorrect coding, faulty tabulation and inappropriate statistical analysis done by the measurer or the researcher, errors may creep into the measurement.

(iii) *Situation*: Situational factors also contribute to errors in measurement. Any situation that puts unnecessary strain on the respondent, tends to introduce errors in measurement. An interview is one example of such a situation. Apart from this, if the person feels that the situation does not protect anonymity, it also introduces error in measurement.

(iv) *Test instrument*: Errors may be introduced in the measurement due to poor psychometric qualities of the test and defective measuring instrument. Psychological tests having poor reliability and validity may result in measurement errors.

To avoid the errors resulting from the above-mentioned sources, it is essential that the researcher should try to eliminate or neutralise these possible errors as for as possible.

DIFFERENCE AMONG ASSESSMENT, TESTING AND MEASUREMENT

The terms assessment, testing and measurement are often confused. Therefore, it is essential that these terms should be clearly distinguished. Assessment is a general term that includes any of a variety of procedures used to obtain information about performance. It includes paper-and-pencil tests, as well as extended responses, performance of authentic tasks (such as laboratory experiment), teacher observation, and self-report of the student. The basic goal of assessment is to evaluate a person in terms of current and future functioning. In the process of assessment, behaviours are classified into different categories measured against normative standard. In fact, assessment answers the basic question: *How well does the individual perform?*

Testing is done through systematic procedure for measuring a sample of behaviour by putting a set of questions in a uniform manner. Such systematic procedure is called test. Thus, tests are used in assessment processes. Since tests are a form of assessment, they also answer the related question: *How well does the individual perform either in comparison with others or in comparison with performance of any task?* However, not all assessment techniques are tests. In the strict sense, any assessment technique is called a test only when its procedure for administration, scoring and interpretation are standardized; there is a standardization sample; and there is evidence for its reliability and validity. Many so-called tests can more appropriately be called as *assessment devices* because they don't meet the minimum requirements of a test.

Measurement, as discussed in the beginning of the chapter, is only a process of obtaining a numerical description of the degree to which an individual possesses a particular characteristic. This numerical description is done according to some rules. In fact, measurement answers the basic question: *How much?*

☑ Review Questions

1. Make a distinction between measurement and evaluation.
2. Discuss the different levels of measurement and give appropriate examples.
3. What do you mean by measurement? Discuss the important functions of measurement in education and psychology.
4. Distinguish between psychological measurement and physical measurement. Give relevant examples.
5. Discuss the general problems of measurement and give suitable examples.
6. Explain the nature and characteristics of measurement in psychology. Distinguish between ordinal and interval scales of measurement.
7. What is meant by psychological measurement? Discuss its main requirements.
8. What is meant by psychological measurement? State the general problems related to measurement in psychology.
9. Distinguish between testing and assessment. Discuss the various sources of errors in measurement.
10. Discuss the various properties of scales that distinguish between the different levels of scales of measurement.

✥

2

TEST CONSTRUCTION

CHAPTER PREVIEW

- Meaning of Test in Psychology and Education
- Classification of Test
- Characteristics of a Good Test
  - (i) Objectivity
  - (ii) Reliability
  - (iii) Validity
  - (iv) Norms
  - (v) Practicability/Usability
- General Steps of Test Construction
  - (i) Planning of the Test
  - (ii) Writing Items of the Test
  - (iii) Preliminary Administration of the Test
  - (iv) Reliability of the Find Test
  - (v) Validity of the Find Test
  - (vi) Norms of the Final Test
  - (vii) Preparation of Manual, and Reproduction of the Test
- Uses and Limitations of Psychological Tests and Testings
- Ethical Issues in Psychological Testing
- Steps in Selecting appropriate published Test

MEANING OF TEST IN PSYCHOLOGY AND EDUCATION

According to the dictionary, 'test' is defined as a series of questions on the basis of which some information is sought. In psychology and education, the meaning of test is something more than this. A psychological (or an educational) test is a standardized procedure to measure quantitatively or qualitatively one or more than one aspect of a trait by means of a sample of verbal or nonverbal behaviour. The purpose of a psychological test is twofold. First, it attempts to compare the same individual on two or more than two aspects of a trait; and second, two or more than two persons may be compared on the same trait. Such a measurement may be either quantitative or qualitative. In the words of Bean (1953, 11), a test is "an organized succession of stimuli designed to measure quantitatively or to evaluate qualitatively some mental process, trait or characteristic". Likewise Anastasi and Urbina (1997) have defined a psychological test as "essentially an objective and standardized measure of sample of behaviour". Similarly, Cullari (1998) has said, "A test is a standardized procedure for sampling behaviour and describing it with scores or categories". Kaplan and Saccuzzo (2001) have opined, "A psychological test or educational test is a set of items designed to measure characteristics of human beings that pertain to behaviour". These definitions reveal some important characteristics of a psychological and educational test.

First, test is an organized succession of stimuli, which means that the stimuli (popularly known as items) in the test are organized in a certain sequence and are based upon some principles of test construction. Usually, the items of a test are processed through item analysis and its procedure of administration is standardized to ensure maximum objectivity. In fact, standardized procedures become essential for ensuring that testing procedures remain uniform for different examiners in different situations. The lack of standardization may change not only the character of the test but also its difficulty level which may ultimately reduce the validity of the test.

Second, both quantitative and qualitative measurements are possible through psychological and educational tests. The reading ability of a child may be quantitatively measured with the help of a test specially designed for the purpose. His reading-ability score may be evaluated (or qualitatively measured) with respect to the average performance of the reading ability of the other children of his age or class. Thus, a test provides both quantitative and qualitative measurement of a trait.

Third, a psychological test is based upon a limited sample of behaviour. What does this mean? Obviously, by this it is meant that any psychological test or educational test does not assess the totality of a person's behaviour. Rather it is focussed on the limited effect of that behaviour. For example, when testing the vocabulary of a person, the test constructor must settle for a sample of 40 to 50 words and predict for that person's word knowledge from this limited sample. In reality, the totality of the person's word knowledge might be poorer or stronger than the 40- or 50-word vocabulary test. Obviously, then, the implication of the test-as-sample concept is that the test score invariably contains some degree of error. Such measurement errors can be minimized by means of a careful test design, but it can never be fully eliminated.

Fourth, psychological tests usually provide scores or categories which are, subsequently, interpreted with reference to a standardization sample. The standardization sample should be representative of the population for whom the test is meant so that it may be possible to evaluate each person's test score or results in comparison to the reference group. For example, knowing that a college student scored 120 on a test of abstract reasoning conveys little meaning. But if we know that the average score for a college student was 110 and that only 2 per cent of these students scored 120 or above, we have a definite basis for making a test prediction, that is we can say that the examinee has a good prospect in college. This point, in itself, illustrates that it is not the result per se that is valuable—rather what is valuable is the test result that is signified in relation to the nontest behaviours, which is of primary interest.

Some psychological tests are *norm-referenced* tests which means that results from them are interpreted with reference to the average performance of the standardization sample whereas some psychological tests are *criterion-referenced*, which are, in fact, used to determine where an examinee stands with reference to a tightly defined criterion or educational objective. On such tests, the comparison is done with an objective standard rather than with the performance of other examinees. For example, results of the criterion-referenced arithmetic test might state that a student does the simple arithmetic works like addition, subtraction, multiplication, division of three digits with 80% accuracy whereas the goal of the school system is 90%. Thus, here, the performance of other students is irrelevant and what is relevant is whether the student meets the accepted criterion.

Finally, for getting the meaning of a psychological test or educational test clear, it is also essential that we must make a distinction between *testing* and *assessment*. In fact, testing is a more limited venture which primarily consists of administering, scoring and interpreting the test scores. Assessment, on the other hand, is a more comprehensive and wider term that includes the entire process of compiling and synthesizing the information to make a prediction about the person. (Also see Chapter 1)

CLASSIFICATION OF TESTS

Psychologists and educators have taken pains over classifying tests from the point of view of different criteria. A brief introduction to this classification may be presented as follows:

1. On the basis of the criterion of administrative conditions

Tests have been classified into two types on the basis of administrative conditions—**individual tests** and **group tests**. Individual tests are those tests that are administered to one person at a time. Kohs Block Design Test is an example of an individual test. Individual tests are often used by school psychologists and counsellors to motivate children and to observe how they respond. Some individually administered tests are given orally, and they require the constant attention of the examiner. Individual tests, in general, have two limitations, i.e., such tests are time-consuming and require the services of trained and experienced examiners. As such, these tests are used only when a crucial decision is necessary.

Group tests are tests which can be used among more than one person or in a group at a time. Bell Adjustment Inventory is an example of the group test. Besides assessing adjustment, group tests are adequate for measuring cognitive skills to survey the achievements, strengths and weaknesses of the students in the classroom, etc.

2. On the basis of the criterion of scoring

Scoring is one of the vital parts of a test. Based upon this criterion, tests are classified into two types—**objective test** and **subjective test**. Objective tests are those whose items are scored by competent examiners or observers in such a way that no scope for subjective judgement or opinion exists and thus, the scoring remains unambiguous. Tests having multiple-choice, true-false and matching items are usually called objective tests. In such items, the problem as well as its answer is given along with the distractor. The problem is known as the **stem** of the item. A distractor answer is one which is similar to the correct answer but is not actually the correct one. Such tests are also known as **new-type tests** or **limited-answer tests**.

Subjective tests are tests whose items are scored by the competent examiners or observers in a way in which there exists some scope for subjective judgement and opinion. As a consequence, some elements of vagueness and ambiguity remain in their scoring. These are also called **essay tests**. Such tests are intended to assess an examinee's ability to organize a comprehensive answer, recall and select important information, and present the same logically and effectively. Since in these tests the examinee is free to write and organize the answer, they are also known as **free-answer tests**. The following items illustrate the nature of an essay test:

(i) Discuss the role of past experience in perception.

(ii) What are the major goals of education? To what extent have these goals been achieved in India?

When an examinee answers these questions or other similar items, he usually selects, recalls and organizes his experiences in the manner he likes.

3. On the basis of the criterion of time limit in producing the response

Another way of classifying tests is whether they emphasize time limit or not. On the basis of this criterion, the tests are classified into **power tests** and **speed tests**. A power test is one which has a generous time limit so that most examinees are able to attempt every item. Usually such tests have items which are generally arranged in increasing order of difficulties. Most of the intelligence tests and aptitude tests belong to the category of power tests. In fact, power tests demonstrate how much knowledge or information the examinees have.

Speed tests are those that have severe time limits but the items are comparatively easy and the difficulties involved therein are more or less of the same degree. Here, very few examinees are supposed to make errors. Speed tests, generally, reveal how rapidly, i.e., with what speed the examinees can respond within a given time limit. Most of the clerical aptitude tests belong to this very category.

In fact, whether a test is a power test or a speed test depends, in part, on the nature of the examinees for whom it is meant. An arithmetical test for class VII students might emphasize speed if it contained items that were easier for them, but the same test could be a power test for class III or IV students or for less-prepared students. Today, a pure power test or pure speed test is rare, rather a mixture of the two is common.

4. On the basis of the criterion of the nature or contents of items

A test may be classified on the basis of the nature of the items or the contents used therein. Important types of tests based on this criterion are:

(i) **Verbal test**
(ii) **Nonverbal test**
(iii) **Performance test**
(iv) **Nonlanguage test**

(i) A verbal test is one whose items emphasize reading, writing and oral expression as the primary mode of communication. Herein, instructions are printed or written. These are read by the examinees and, accordingly, items are answered. Jalota Group General Intelligence Test and Mehta Group Test of Intelligence are some common examples. Verbal tests are also called paper–pencil tests because the examinee has to write on a piece of paper while answering the test items.

(ii) Nonverbal tests are those that emphasize but don't altogether eliminate the role of language by using symbolic materials like pictures, figures, etc. Such tests use the language in instruction but in items, they don't use language. Test items present the problem with the help of figures and symbols. Nonverbal tests are commonly used with young children as an attempt to assess the nonverbal aspects of intelligence such as spatial perception. Raven Progressive Matrices is a good example of nonverbal test.

(iii) Performance tests are those that require the examinees to perform a task rather than answer some questions. Such tests prohibit the use of language in items. Occasionally, oral language is used to give instruction, or, the instruction may also be given through gesture and pantomime. Different kinds of performance tests are available. Some tests require examinees to assemble a puzzle, place pictures in a correct sequence, place pages in the boards as rapidly as possible, point to a missing part of the picture, etc. One feature of performance tests is that they are usually administered individually so that the examiner can count the errors committed by the examinee or the student and can assess how long it takes him to complete the given task. Whatever may be the types of performance test, the common feature of all performance tests is their emphasis on the examinee's ability to perform a task rather than answer some questions.

(iv) Nonlanguage tests are those which don't depend upon any form of written, spoken or reading communication. Such tests remain completely independent of the ability to use language in any way. Instructions are usually given through gestures or pantomime and the examinees respond by pointing at or manipulating objects such as pictures, blocks, puzzles, etc. Such tests are usually administered to those persons or children who can't communicate in any form of ordinary language.

5. On the basis of the criterion of purpose or objective

Tests are also classified in terms of their objectives or purposes. Based upon this criterion, tests are usually classified as intelligence tests, aptitude tests, personality tests, neuropsychological tests and achievement tests. Intelligence tests intend to assess intelligence of the examinees. Aptitude tests assess potentials or aptitudes of the persons. Personality tests assess traits, adjustments, interests, values, etc., of the persons. Neuropsychological tests are the tests, which are used in the assessment of persons with known or suspected brain dysfunctioning. Achievement tests assess what the persons have acquired in the given area as a function of some training or learning.

6. On the basis of the criterion of standardization

Tests are also classified on the basis of standardization. Based upon this criterion, tests are classified into standardized tests and teacher-made tests. Standardized tests are those which have been subjected to the procedure of standardization. However, the meaning of the term 'standardization' is controversial and includes at least the following conditions:

(i) The first condition for standardization is that there must be a standard manner of giving instructions so that uniformity can be maintained in the evaluation of all those who take the test.

(ii) The second condition for standardization is that there must be uniformity of scoring and an index of fairness of correct answer through the procedure of item analysis should be available.

(iii) The third condition is that reliability and validity of the test must be established and the individuals for whom the test is intended should be explicitly mentioned.

(iv) The fourth condition, a controversial one, is that a standardized test should have norms. However, according to Cronbach (1970, 27), a test even without norms may be called a standardized test. But the majority of psychologists favour the idea that a standardized test should have norms as well.

By way of summarizing the meaning of a standardized test, it can be said that standardized tests, constructed by test specialists, are standardized in the sense that they have been administered and scored under standard and uniform testing conditions so that the results obtained from different samples may legitimately be compared. Items of standardized tests are fixed and not modifiable.

Teacher-made tests are those that are constructed by teachers for use largely within their classrooms. The effectiveness of such tests depends upon the skill of the teacher and his knowledge of test construction. Items may come from any area of curriculum and they may be modified according to the will of the teacher. Rules for administration and scoring are determined by the teacher. Such tests are largely evaluated by the teachers themselves and no particular norms are provided; however, they may be developed by the teacher for his own class.

Thus, we find that tests have been classified in terms of various criteria. These tests are used for a variety of purposes.

CHARACTERISTICS OF A GOOD TEST

For a test to be scientifically sound, it must possess the following characteristics.

Objectivity

A test must have the trait of objectivity, i.e., it must be free from the subjective element so that there is complete interpersonal agreement among experts regarding the meaning of the items and scoring of the test. Obviously, objectivity here relates to two aspects of the test—objectivity of the items and objectivity of the scoring system. By objectivity of items is meant that the items should be phrased in such a manner that they are interpreted in exactly the same way by all those who take the test. For ensuring objectivity of items, items must have uniformity of order of presentation (that is, either ascending or descending order). By objectivity of scoring is meant that the scoring method of the test should be a standard one so that complete uniformity can be maintained wher the test is scored by different experts at different times.

Reliability

A test must also be reliable. Reliability here refers to self-correlation of the test. It shows the extent to which the results obtained are consistent when the test is administered once or more than once on the same sample with a reasonable time gap. Consistency in results obtained in a single administration is the index of internal consistency of the test and consistency in results obtained upon testing and retesting is an index of temporal consistency. Reliability, thus, includes both internal consistency as well as temporal consistency. For a test to be called sound, it must be reliable because reliability indicates the extent to which the scores obtained in the test are free from such internal defects of standardization which are likely to produce errors of measurement.

Validity

Validity is another prerequisite for a test to be sound. Validity indicates the extent to which the test measures what it intends to measure, when compared with some outside independent criterion. In other words, it is the correlation of the test with some outside criterion. The criterion should be an independent one and should be regarded as the best index of trait or ability being measured by the test. Generally, validity of the test is dependent upon the reliability because a test which yields inconsistent results (poor reliability) is ordinarily not expected to correlate with some outside independent criterion.

Norms

A test must also be guided by certain norms. Norms refer to the average performance of a representative sample on a given test. There are four common types of norms—age norms, grade norms, percentile norms and standard score norms. Depending upon the purpose and use, a test constructor prepares any of these norms for his test. Norms help in interpretation of the scores. In the absence of norms, no meaning can be added to the score obtained on the test.

Practicability / Usability

A test must also be practicable/usable from the point of view of the time taken in its completion, length, scoring, etc. In other words, the test should not be lengthy and the scoring method must not be difficult nor one which can only be done by highly specialized persons. In addition, the test should be economical from the point of view of money also.

GENERAL STEPS OF TEST CONSTRUCTION

Before the real work of test construction begins, certain broad decisions are taken by the investigator. These preliminary decisions have far-reaching consequences. It is at this preliminary stage that the test constructor outlines the major objectives of the test in general terms, and specifies the populations for whom the test is intended. He also indicates the possible conditions under which the test can be used and its important uses. For example, a test constructor may decide to construct an intelligence test meant for students of the tenth grade broadly aiming at diagnosing the manipulative and organizational ability of the pupils. Having decided the above preliminary things, he must go ahead with the following steps.

1. Planning of the test
2. Writing items of the test
3. Preliminary administration (or the experimental try-out) of the test
4. Reliability of the final test
5. Validity of the final test
6. Norms of the final test
7. Preparation of manual and reproduction of the test

Planning of the Test

The first step in the construction of a test is careful planning. At this stage, the test constructor specifies the broad and specific objectives of the test in clear terms. He decides upon the nature of the content or items to be included, the type of instructions to be included, the method of sampling, a detailed arrangement for the preliminary administration and the final administration, a probable length and time limit for the completion of the test, probable statistical methods to be adopted, etc. Planning also includes the total number of reproductions of the test to be made and a preparation of a manual.

Writing of the Test Items

The second step in test construction is the preparation of the items of the test. According to Bean (1953, 15), an item is defined as "a single question or task that is not often broken down into any smaller units." Item writing starts with the planning done earlier. If the test constructor decides to prepare an essay test, the essay items are written down. However, if he decides to construct an objective test, he writes down the objective items such as the alternative reponse item, matching item, multiple-choice item, completion item, short-answer item, pictorial form of item, etc. Depending upon the purpose, he decides to write any of these objective types of items.

Item writing is essentially a creative art. There are no set rules to guide and guarantee writing of good items. A lot depends upon the item writer's intuition, imagination, experience, practice and ingenuity. However, there are some essential prerequisites, which must be met if the item writer wants to write good and appropriate items. These requirements are enumerated as follows:

1. The item writer must have a thorough knowledge and complete mastery of the *subject matter*. In other words, he must be fully acquainted with all facts, principles, misconceptions, fallacies in a particular field so that he may be able to write good and appropriate items.

2. The item writer must be fully aware of those *persons* for whom the test is meant. He must be aware of the intelligence level of those persons so that he may manipulate the difficulty level of the items for proper adjustment with their ability level. He must also be able to avoid irrelevant clues to the correct responses.

3. The item writer must be familiar with different types of items along with their *advantages* and disadvantages. He must also be aware of the characteristics of good items and the common probable errors in writing items.

4. The item writer must have a large *vocabulary*. He must know the different meanings of a word so that confusion in writing the items may be avoided. He must be able to convey the meaning of the items in the simplest possible language.

5. After writing down the items, they must be submitted to a group of subject experts for their criticisms and suggestions, following which the items must then be duly modified.

The item writer must also cultivate a rich source of ideas for items. This is because ideas are not produced in the mind automatically, rather, they require certain factors or stimuli. The common source of such factors are textbooks, journals, discussions, questions for interview, course outlines, and other instructional materials. After the items have been written down, they are reviewed by some experts or by the item writer himself and then arranged in the order in which they are to appear in the final test. Generally, items are arranged in an increasing order of difficulty, and those having the same form (say, alternative form, matching, multiple-choice, etc.) and dealing with the same contents are placed together.

Preliminary Administration (or the Experimental Try-out) of the Test

When the items have been written down and modified in the light of the suggestions and criticisms given by the experts, the test is said to be ready for its experimental try-out. The purpose of experimental try-out or preliminary administration of the test is manifold. According to Conrad

(1951), the main purpose of the experimental try-out of any psychological and educational test is as given below:

1. Finding out the major weaknesses, omissions, ambiguities and inadequacies of the items. In other words, try-out helps in identifying the ambiguous and indeterminate items, nonfunctioning distractors in multiple-choice items, very difficult or very easy items, and the like.

2. Determining the difficulty values of each item which, in turn, helps in selecting items for their even and proper distribution in the final form.

3. Determining the validity of each individual item. The experimental try-out helps in determining the discriminatory power of each individual item. The discriminatory power here refers to the extent to which any given item discriminates successfully between those who possess the trait in larger amounts and those who possess the same trait in the least amount.

4. Determining a reasonable time limit of the test.

5. Determining the appropriate length of the test. In other words, it helps in determining the number of items to be included in the final form.

6. Determining the intercorrelations of items so that overlapping can be avoided.

7. Identifying any weakness and vagueness in directions or instructions of the test as well as in the fore-exercises or sample questions of the test.

For achieving these aims of experimental try-out, Conrad (1951) recommended at least three preliminary administrations of the test. The aim of the first administration is to detect any gross defects, ambiguities, and omissions in items and instructions. For the first administration, the number of examinees should not be less than 100. He refers to the first try-out as the "Pre-try-out". The aim of the second preliminary administration is to provide data for item analysis, and for this the number of examinees should be around 400. Conrad calls this second try-out "the try-out proper". The sample for this must be similar to those for whom the test is intended. *Item analysis is a technique of selecting discriminating items for the final composition of the test. It aims at obtaining three kinds of information regarding the items: (i) the difficulty value of the item, (ii) the discrimination index of the item, and (iii) the effectiveness of distractors.* (A detailed discussion of item analysis would appear in Chapter 4.) The third preliminary administration is carried out to detect any minor defect that may not have been detected by the first two preliminary administrations. Conrad calls this third try-out the "final trial administration". At this stage, the items are selected after item analysis and they constitute the test in the final form. The "final trial administration" indicates how effective the test will really be when it would be administered on the sample for which it is really intended. Thus the third preliminary administration would be a kind of 'dress rehearsal' providing a sort of final check on the procedure of administration of the test and its time limit. After the final trial administration is over, no material change is ordinarily to be included in the test.

Although the procedures recommended by Conrad for preliminary administrations have been widely appreciated, it has not been followed as a fixed rule. As a matter of fact, the number of preliminary administrations and the number of the examinees for each administration have widely varied depending upon the nature of the test as well as upon the purpose of the test.

Reliability of the Final Test

When on the basis of the experimental or empirical try-out the test is finally composed of the selected items, the final test is again administered on a fresh sample in order to compute the reliability coefficient. The size of the sample for this purpose should not be less than 100. Reliability is the self-correlation of the test and it indicates the consistency of the scores in the test. There are three common ways of calculating reliability coefficient, namely test-retest method, split-half method, and the equivalent-form method. Besides these, the Kuder Richardson formulas and the Rulan formula are also used in computing the reliability coefficient of the test. (A detailed discussion of these methods appears in Chapter 5.)

Validity of the Final Test

Validity refers to what the test measures and how well it measures. If a test measures a trait that it intends to measure well, we say that the test is a valid one. After estimating the reliability coefficient of the test, the test constructor validates the test against some outside independent criteria by comparing the test with the criteria. Thus, validity may also be defined as the correlation of the test with some outside independent criteria. Validity should be computed from the data obtained from the samples other than those used in item analysis. This procedure is known as *cross-validation*. There are three main types of validity: content validity, construct validity and criterion-related validity. The usual statistical techniques employed in computing validity coefficients are Pearsonian *r*, biserial *r*, pointbiserial *r*, chi square, phi-coefficient, etc. The abac tables have also been prepared by Flanagan for directly reading the values of biserial *r*, pointbiserial *r* and phi-coefficient when the proportions of those passing an item in the lower group and upper group are known. (A detailed discussion regarding these types of validity and statistical techniques employed appears in Chapter 6.)

Norms of the Final Test

Finally, the test constructor also prepares norms of the test. Norms are defined as the average performance or score of a large sample representative of a specified population. Norms are prepared to meaningfully interpret the scores obtained on the test for, as we know, the obtained scores on the test themselves convey no meaning regarding the ability or trait being measured. But when these are compared with the norms, a meaningful inference can immediately be drawn. The common types of norms are the age norms, the grade norms, the percentile norms, and the standard score norms. All these types of norms are not suited to all types of tests. Keeping in view the purpose and type of test, the test constructor develops a suitable norm for the test. The preliminary considerations in developing norms are that the sample must be representative of the true population; it must be randomly selected; and it should preferably represent a cross-section of the population. (A detailed discussion of the different types of norms appears in Chapter 7.)

Preparation of Manual and Reproduction of the Test

The last step in test construction is the preparation of a manual of the test. In the manual the test constructor reports the psychometric properties of the test, norms and references. This gives a clear indication regarding the procedures of the test administration, the scoring methods and time limits, if any, of the test. It also includes instructions as well as the details of arrangement of materials, that is, whether items have been arranged in random order or in any other order. In general, the test manual should yield information about the standardization sample, reliability, validity, scoring as well as practical considerations. The test constructor, after seeing the importance and requirement of the test, finally orders for printing of the test and the manual.

USES AND LIMITATIONS OF PSYCHOLOGICAL TESTS AND TESTINGS

Psychological tests are widely used for many purposes. It is very convenient to distinguish the following five uses of tests:

1. **In classification:** Psychological tests are popularly used in making classification of persons, that is, for assigning the persons to one category rather than to another one. There are different types of classification, each one giving emphasis upon a particular purpose in assigning persons to categories. Important types of categories are *placement, screening, certification* and *selection*, where psychological tests play a significant role in each of these types.

Placement refers to the sorting of persons into appropriate programmes according to their needs or skills. With the help of the appropriate psychological tests, teachers often in class enroll some of the students into science faculty and some of them into social science faculty. They enroll some students for mathematics, physics and chemistry programme whereas others are enrolled for geography, history and political science programme. Without the help of psychological tests, it is not possible to do such placement.

Screening refers to the procedures of identification of persons with special characteristics or needs. With the help of psychological tests, psychometricians often screen persons into creative persons and persons having exceptional talent in abstract reasoning. They administer the tests and on the basis of the scores obtained, are able to screen them in desired categories.

Certification and *selection* are done with the help of psychological tests. Certification implies that an individual has at least a minimum proficiency in some discipline or activity. When a person passes a certification examination, it automatically confers some privileges. For example, when a driver passes a driving examination, he gets a license. This illustrates the process of certification. Selection is very much similar to certification because it also confers some privileges on the part of the persons who have been selected. These tasks are well accomplished with the help of psychological tests. The persons who are selected on the basis of the test scores, for example, get admission into a certain course or gain employment in an organization.

2. **In diagnosis and planning for treatment:** Psychological tests play a significant role in making diagnosis and in planning for treatment. Diagnosis means determining the nature of a person's abnormal behaviour and classifying the behaviour pattern within an accepted system. Intelligence tests are considered important for diagnosis of mentally retarded children. Likewise, through some psychological tests, a diagnosis of learning disability can easily be done with the help of MMPI, a clinical psychologist readily diagnoses persons with pathological traits. A proper diagnostic programme not only provides assignment of a label, but also the choice for treatment. When a child is diagnosed as mentally retarded or having learning disability, a planning for his treatment is accordingly done so that the maximum help can be rendered.

3. **In self-knowledge:** Psychological tests are also useful in providing self-knowledge to the test takers to the extent that such knowledge tends to change their career path. Every administration of a psychological test gives a feedback to the test takers regarding the level of trait/ability being assessed. As a consequence, they bring a change in the desired direction and mould their path for betterment.

4. **In evaluation of programmes:** Psychological tests are often used in evaluation of various types of educational and social programmes. In schools and colleges, different types of programmes for betterment of academic achievement are carried out and the persons want to know about its impact. Such impacts are easily assessed with the help of various types of achievement and intelligence tests. Likewise, people in general and political parties in particular want to assess the outcome of a social programme carried out for the purpose of say, verifying the IQ levels of disadvantaged group. This is also done with the help of various types of psychological tests.

5. **In theoretical and applied branches of behavioural research:** Psychological tests are very useful in research. They are frequently used in both theoretical and applied researches. With the help of such tests, psychologists frequently investigate theoretical matters that have no immediate or obvious practical applications. Here, we can cite the example of Witkin (1949) who for analyzing perceptual field dependence, developed the *tilting-room-tilting-chair test* (TRTC). In fact, TRTC encouraged a good deal of research on personality development but was seldom applied to any practical problems of testing. Take the example for an applied field. Suppose neuropsychologists wish to test the hypothesis that low level of lead absorption produces behavioural deficit in children. This hypothesis can be easily tested by examining lead-burdened children and normal children with the help of psychological tests. With the help of various types of psychological tests, it has been reported that low-level lead absorption in children produces decrement in IQ, impairment in reaction time and increase in undesirable classroom behaviour. This automatically shows that psychological tests are useful in applied areas too and there should not be any debate about the validity of testing-based research findings.

Thus, we see that psychological tests are used in varieties of fields. Often students remain curious to know from where to obtain information about psychological tests. In particular, here, it can be said that information about psychological tests is available from five sources: *Publisher's Catalogues, Reference books, Journals, Databases* and *Test manuals*. The best single reference source of information is Mental Measurements Year book (MMY) which is reviewed time and again to incorporate the new tests. Information can also be had from the catalogues of major test publishers. Appendix C provides the names of foreign and Indian Test publishers. Information about tests can also be had from various journals. Appendix D provides the list of important journals in the field of behavioural researches.

Despite its various uses, psychological tests are not without limitations. Some of the important limitations are as under:

(i) *Psychological tests represent an invasion of privacy*: Psychological tests may be invasion of privacy if they are used without the permission of the testees to obtain personal and sensitive information.

(ii) *Psychological tests permanently categorize the persons*: On the basis of the performance of psychological tests, the testees or examinees, are given certain categories like mentally retarded, gifted, brain-damaged, etc., and the authority behaves accordingly disregarding evidence of any further change. This has a serious implication for the examinees. The examinees can definitely change and great care should be taken in the interpretation and use of the test results.

(iii) *Psychological tests measure only limited and beneficial aspects of behaviour*: It is said that psychological tests cannot measure the most important human traits. They force the examinees to take decisions based on superficial and relatively unimportant criteria.

(iv) *Psychological tests create anxiety*: Generally, it has been reported that when the assessment is to be done through psychological tests, the examinees feel anxious and this anxiety affects their performances. However, the examinees who are familiar with specific types of tests are less anxious than those who are familiar with the test contents (Sax, 1974).

(v) *Psychological tests penalize bright and creative examinees*: Psychological tests are insensitive to atypical and creative responses. Such responses are not given much credit, thus discriminating against the talented examinees.

Thus psychological tests have some limitations.

ETHICAL ISSUES IN PSYCHOLOGICAL TESTING

As we know, psychological testing refers to all the possible uses, applications and underlying important concepts of psychological and educational tests. To maintain its proper uses and applications, the American Psychological Association (APA) officially adopted a set of standards and rules in 1953 which have undergone continual review and refinement. The current version, called *Ethical Principles of Psychologists and Code of conduct* (APA 1992), consists of a preamble and six general principles which guide psychologists towards the highest ideals in their profession. In addition, it also provides eight ethical standards with enforceable rules for psychologists who are working in different contexts. The *APA Committee on Psychological Tests and Assessment* (CPTA) is especially designed for considering problems regarding sound testing and assessment practices and for providing various types of technical advices.

The major ethical or moral issues relating to psychological testing can be described under the following five headings:

1. *Issues of human rights*: Today the field of psychological testing has been heavily influenced by recognition of various types of human rights. Among these rights is the *right to not to be tested*. In fact, persons who don't want to subject themselves to testing, should not and ethically can't be forced to accept this. Moreover, individuals who finally decide to subject themselves to testing, have rights to know their test scores, their interpretations as well as the basis of any decisions that affect their lives. In the name of guarding the security of tests, test interpreters cannot deprive the test takers (or subjects) from the right to know the basis of detrimental or adverse decisions. Likewise, these days other human rights such as the right to know who will have access to the data of psychological testing, and the right to confidentiality of test results are being popularly discussed. Test interpreters have an ethical obligation to provide protection to these human rights whereas potential test takers are responsible for demanding their rights. Such awareness of human rights today is casting a very important influence on psychological testing and also shaping its future.

2. *Issue of labelling*: On the basis of psychological testing, a person is given a certain label or diagnosed as having a certain psychiatric disorder. This labeling has many harmful effects. For example, suppose a person has been diagnosed with *chronic schizophrenia*, which in fact, has a little chance of being cured. In fact, labelling someone a chronic schizophrenic may be a self-fulfilling prophecy. Since this disorder is incurable, nothing can be done and when nothing can be done why should one bother to provide help to such a person. Because no help is given, the person remains a chronic case. Thus, labelling can stigmatize a person for life and it also affects one's access to help. Such a labelling creates additional problems. When a person is labelled as schizophrenic, it automatically implies that he is not responsible for it because schizophrenia is a disease or illness and nobody can be blamed for becoming ill. Therefore, such labelling will make him passive and will leave no incentive to alter the negative conditions surrounding his life. Therefore, labelling will not only stigmatize persons but it will also lower tolerance for stress and make treatment difficult. In view of these potential negative effects and dangers of labelling, a person should have the right to not be labelled.

3. *Issues of invasion of privacy*: When people respond to items of psychological tests, they have little idea of what is being revealed by their responses but somehow, they feel that their privacy has been invaded. Such a feeling is definitely detrimental for people.

In fact, there are two sides to this issue. Dahlstrom (1969) investigated this issue in detail and pointed out two related aspects of this issue. He has pointed out that this issue of invasion of privacy is based on serious misunderstandings. In fact, psychological tests have very limited and pinpointed aim and they *can't* invade the privacy of the persons. Another aspect of this issue, again pointed out by Dahlstrom (1969), is the ambiguous nature of the notion of invasion of privacy itself. In reality psychologists don't consider it wrong, evil or even detrimental to find out or collect information about the person. The person's privacy is invaded when most information is used inappropriately or wrongly. Psychologists are ethically and even legally bound to maintain confidentiality and they don't reveal any more information about a person than is necessary to accomplish the purpose for which the testing was started. In fact, the ethical code of APA (1992) has included *confidentiality*, which obviously dictates that personal information obtained by the psychologist from any source is communicated to others only with the person's consent. Exception to this exists in only those circumstances in which withholding information may cause danger to the person or society as well as cases that have subpoenaed records.

4. *Issues of divided loyalties*: This is one of the vital issues of psychological testing and was first pointed out by Jackson and Messick (1967) and still today remains a central problem in the field of psychological testing. In fact, divided loyalties is today a major dilemma for psychologists who use the test in different fields such as industry, schools, clinics, government, military, and so on. A psychologist has to face a conflict, which arises when the individual's welfare is put at odds on the one hand and that of the institution that employs the psychologist on the other. For example, suppose a psychologist working for an industrial firm to identify individuals who are accident prone, has the responsibility towards the institution to identify such persons as well as the responsibility to protect the rights and welfare of the persons seeking employment. Here, the psychologist's loyality stands divided. Likewise, a psychologist has to maintain test security at any cost but he must also not violate the person's right to know the basis of an adverse decision. However, if this basis is explained to one person, this information may go to the persons with the same problem, who then rightly decide to outsmart the test. Again, the psychologist is trapped with two opposing principles.

5. *Responsibility of test constructors and test users*: Ethical issues also put some responsibilities on test constructors, or developers, and test users. In fact, the test constructor is responsible for providing all the necessary information. Latest standards for test use state that the test constructors must provide a test manual which may clearly state the appropriate use of the test, including data relating to reliability, validity, and norms, clearly specify about the scoring and administration standards (AERA*, APA & NCME* 1999).

There are some responsibilities which lie with test users too. According to APA (1974), almost any test can be useful if it is used in the right circumstances but even the best test can hurt the subject if it is used inappropriately. To minimize such potential damage, APA (1974) makes uses of the tests responsible for knowing the reason for using the test, as well as the consequences of using the test. It also makes test users to maximize the test's effectiveness and to minimize unfairness, if any. In other words, test users must possess sufficient and adequate knowledge to understand the basic principles underlying the test construction and supporting research of any test they administer. They must also be aware of the psychometric qualities of the test being used as well as the relevant literature. At any cost, a test user cannot claim ignorance. The test user is responsible for finding out relevant and pertinent information using any test.

These ethical issues, in fact, put psychologists on vigil so that they must guard both the interest of the test as well as the welfare of the test takers.

STEPS IN SELECTING APPROPRIATE PUBLISHED TESTS

As we know, published tests play a significant role in research as well as in any educational programme of the school. Therefore, they need to be selected with utmost care. When published tests are carefully selected and appropriately used, they make valuable contribution to the research outcomes. On the other hand, when they are selected hastily or casually, they seldom make any significant contribution because they provide inadequate or inappropriate information. The following sequence of *steps* must be followed in selecting a published test:

1. *Defining the needs of testing:* The first step in selecting published tests is to define specifically the purpose of testing and the type of information being sought through testing. The purpose of the testing must be clearly defined and the test should match with the purpose. In any given content area, there are numerous tests and each measures somewhat different aspects of knowledge and skill. To make a proper and appropriate solution, the researcher must first identify the objectives and specific outcomes of the research programme. This is, in fact, necessary in choosing relevant tests, whether selecting a single test or a battery of tests for research or testing programmes.

2. *Narrowing the choice:* The need for published tests should not be considered in isolation, rather in relation to the total measurement programme. In fact, this makes it possible to choose tests that supplement and complement the other means of assessment. For example, if a scholastic aptitude test is to be used for diagnosing students needing extra help in learning to read, it is desirable to replace the aptitude test with a diagnostic reading test. There are also other factors that help narrow the choice. For example, if the test user (or researcher) lacks the adequate experience to administer individual test, only the group tests should be considered. Likewise, if the tests are to be administered by such teachers who lack experience in test administration, tests

---

* AERA stands for **American Educational Research Association** and NCME stands for **National Council on Measurement in Education**.

with simple directions should be the appropriate choice. Similarly, if the same type of achievement test is desired for both elementary and high school levels, only those achievement tests, which are suitable for all grade levels should be given due consideration.

3. *Locating suitable and appropriate tests:* When the researchers have reduced the list of tests to a reasonable number, specimen sets should be obtained and examined so that the test manuals as well as the items of the test may be evaluated properly.

4. *Reviewing test materials:* A test manual provides several types of information for judging the appropriateness and related technical qualities of the test. A good test manual, among others, provide the following types of information:

(i) Major uses for which test is recommended
(ii) Qualifications essential for administering and interpreting the test
(iii) Evidence for reliability of recommended uses
(iv) Evidence for validity of recommended uses
(v) Directions for administering and scoring the test
(vi) Adequate norms for interpreting the scores.

Besides reviewing the test manuals, it is also wise to study carefully the individual test items and the best method of doing this is to try to answer each item as if you were taking the test. Although this is a time-consuming process, there is no other more better means of determining how appropriate a test is for measuring the knowledge or skill.

5. *Using a test evaluation form:* At the last step, the researcher should use a test evaluation form, which makes it easier to gather information about a specific test. In fact, test evaluation form guarantees that the pertinent information will not be overlooked. Such form also provides for a summary comparison of each test's strengths and weaknesses. A specimen of test evaluation form is presented below.

Test Evaluation Form

Title of Test:	Author(s):
Publisher:	Purpose of Test:
Publication date:	Forms:
Types of scoring:	Administration time:
Technical Features:	
• *Reliability:* Nature of evidence (internal consistency, external consistency, standard error of measurement, etc.) • *Validity:* Nature of evidence • *Norms:* Type and appropriateness to local situation	
Practical Features	*General Evaluation*
• Ease of scoring and interpretation • Ease of administration (Procedure and timing) • Adequacy of test manual and related materials	• Summary of strengths • Summary of weakness • Comments of reviewers • Recommendation regarding local use

☑ Review Questions

1. Give the meaning of a test in psychology. Point out the major characteristics of a good psychological test.
2. Make a distinction between a teacher-made test and a standardized test.
3. Outline a plan for classifying a psychological and an educational test.
4. Describe, with examples, the general steps in construction of a psychological test.
5. What is a psychological test? Describe the essential characteristics of a good psychological test.
6. Discuss the nature of standardization of a test. Explain the different aspects taken into consideration for the standardization of a test.
7. Discuss the important uses and limitations of psychological tests.
8. Discuss the important ethical issues involved in psychological testing.
9. Discuss the major steps in selecting appropriate published test.

◆

3

ITEM WRITING

CHAPTER PREVIEW

- Meaning and Types of Items
  - (i) Essay Item
  - (ii) Objective Item
- Difference between Essay-Type Tests and Objective-Type Tests
- General Guidelines for Item Writing
- General Methods of Scoring Objective-Test Items
  - (i) Overlay-key Method
  - (ii) Strip-key Method

MEANING AND TYPES OF ITEMS

For a test to be sound, the items must be good. Bean (1953,15) defines item as "a single task or question that usually cannot be broken down into any smaller units". An arithmetical problem may be an item; a manipulative task may be an item; a mechanical puzzle may be an item; and likewise, sleeplessness may also be an item of a test. It is in this sense that Nunnally (1959) referred to items as the '*lowest common denominator of a test*', which is scored.

An item must have the following characteristics:

1. An item should be phrased in such a manner that there is no ambiguity regarding its meaning for both the item writer as well as the examinees who take the test.
2. The item should not be too easy or too difficult.
3. It should have discriminating power, that is, it must clearly distinguish between those who possess a trait and those who do not.
4. It should not be concerned with the trivial aspects of the subject matter, that is, it must only measure the significant aspects of knowledge or understanding.
5. As far as possible, it should not encourage guesswork by the subject.
6. It should not present any difficulty in reading.
7. It should not be such that its meaning is dependent upon another item and/or it can be answered by referring to another item.

The item writer is also frequently faced with the problem of determining the exact number of items. As a matter of fact, there is no hard and fast rule regarding this. Previous studies have shown that the number of items is usually linked with the desired level of reliability coefficient of the test. Studies have revealed that usually 25 to 30 dichotomous items are needed to have a reliability coefficient as high as 0.80 whereas 15 to 20 items are needed to reach the same level of reliability when multipoint items are used. These are the minimum numbers of items which should be retained after item analysis. An item writer should always write almost twice the number of items to be retained finally. Thus, if he wants 30 items in the final test, he should write 60 items. In the speed test, the number of items to be written is entirely dependent upon the

intuitive judgement of the test constructor. On the basis of his previous experiences, he decides that a certain number of items can be answered within the given time limit.

An item writer must also have sufficient knowledge regarding the different forms or types of items as well as the merits and demerits of each type. In a broad sense, items are of two types:

1. Essay item or free-answer item
2. Objective item or 'new type' of item

Essay Item or Free-answer Item

An essay item is one in which the examinee relies upon his memory and past associations to answer the questions in a few words only. Since such items can be answered in whatever manner one likes, they are also known as free-answer items. Essay items are most appropriate for measuring higher mental processes which involve the process of synthesis, analysis, evaluation, organization and criticism of the events of the past. Essay tests are thus suitable for measuring traits like critical thinking, originality and the ability to integrate or synthesize or analyze different events. Essay items are of two types: *short-answer type* and *long-answer type or extended-answer essay type*. A short-answer essay item is one where the examinee supplies the answer in one or two lines and is usually concerned with one central concept (Marshall & Hales, 1972). A long-answer essay item is one where the examinee's answer comprises several sentences. Such an item is usually concerned with more than one central concept. The following examples illustrate the two varieties of essay items:

1. Explain the meaning of reliability of an educational test. (Short-answer essay item)
2. Describe the methods of estimating reliability and validity of an educational test. (Long-answer or extended-answer essay item)

Nowadays essay items are frequently used by teachers to measure the achievement of pupils in a classroom. The primary merit of such items is that they encourage the examinee to give a coherent and organized picture of his memory and past associations. In other words, they require the examinees to organize and produce the answer rather than recognize the answer. Hence, a better inference can be drawn regarding the examinee on the basis of these answers. The most serious drawback of such an item, however, is that marking is highly unreliable. It varies from scorer to scorer and sometimes within the same scorer, when he is asked to evaluate the answer at different time intervals. Not only this, scoring of the essay items takes a longer time because of the length of the answer and also because the scorer is required to read each line very carefully. Thus essay items lack scoring economy, too. Despite these weaknesses, essay items are popular measures of educational achievement in the classroom.

Essay test items present two important difficulties. *First,* the factual knowledge may sometimes be confounded with the ability to organize and synthesize an answer. The examinee who knows more fact correctly may write an answer that looks better and superior when the originality may be missing from the answer. *Second,* in comparison to objective items where the examinee is required only to select from alternatives, here in essay items, he is required to think out an answer and compose a written response. Naturally, this takes more time. As a consequence, differences in performance can depend on ability to write quickly in correct form. Not only this, due to the greater time required to compose an answer, very few questions can be asked and this may lower the content validity of the essay test.

To combat these problems, several variations in the testing environment have been proposed by psychologists and educators. Following are the important ones:

(i) *Open-book examination:* In this format, the examinees are permitted to use their textbooks or classnotes for answering the questions. This technique reduces the impact of differences in knowledge. The demerit of this method is that the less able examinees may devote much time in searching material during the test whereas the more able examinees devote much time in composing the answers. Thus, the examinees who devote more time in searching the materials in books or notes may feel that they have less time than if the examination would have been on the pattern of closed-book.

(ii) *Take-home examination:* As its name implies, in this type of examination the examinees don't have time pressure during the test and it allows the examinees to work in a completely unsupervised situation as long as he or she wishes (or has the time to do). One major drawback of this format is that it sacrifices test security. In fact, there is no way to know that the examinee has worked independently. This apprehension becomes still more stronger due to internet information and use of cell phones.

(iii) *Cheat sheets:* Cheat sheet is one variation of open-book examination. Here the examinee is allowed to bring to examination one piece of paper with as many facts, procedures, points or other information as desired. This technique has been found to work well in courses where formulas or several small points or facts constitute a large part of the subject matter. This technique reduces dependence on the textbooks.

(iv) *Study questions:* This technique combines features of both the open-book and take-home examination. Here the examinee is provided with a list of study questions from which examination questions will be set or drawn. For example, the examinee may be provided a list of 10 to 15 questions and told that the examination will contain 4 or 5 of them. This technique has the advantage that the examinees can take as much time as they wish for preparing for the examination and can also use whatever resources are available to them intensively.

In order to write good essay items, the following suggestions must be kept in mind:

1. An essay item must contain explicitly defined problems. Usually, essay items are intended to measure the higher mental processes. As such, it is essential that they contain problems in clear-cut and explicit terms so that every examinee interprets them in more or less the same way. An essay item is not said to be valid if its interpretation varies among examinees.

2. It must contain such problems whose answers are not very wide. In case a student is asked to answer a problem with a larger content area, he may start writing whatever he knows without making any discrimination. In such a situation he may not write about the facts or information needed by the item, thus lowering the validity of the essay item.

3. Essay items must have clear-cut directions or instructions for the examinees. The instruction should indicate the total time to be spent on any particular item, what type of information is required, and the likely weightage to be given to each item so that the examinee may pick up the relative importance of the essay questions and accordingly, adjust the length of the answers.

4. Sufficient time should be allowed in the construction of the essay items. Such items measure the higher mental processes and in order that they actually measure what they intend to measure, it is essential that they are carefully worded and ordered so that they are interpreted in the same way by all.

5. One should start essay item/question with phrases such as 'compare', 'contrast', 'criticize', 'Explain how', 'differentiate', etc. The use of these words or phrases will help present the examinees with task requiring them to select, organize and apply information. Essay items should not be started with 'what', 'when', 'who' or 'list' because such words encourage examinees only to the production of facts.

6. The item writer must clearly know what mental processes he wants to assess before starting to write the item/questions. He must fully understand the kinds of responses that represent the abilities that he wishes to assess. Once the abilities or competencies have been decided, he can select, adopt or create items/questions that will require the examinees to display those abilities.

7. The item writer should use novel material or novel organisation of material in phrasing the questions. Essay items always intend to test the examinees' ability to use their information. For this, it is essential that the item writer must put them in situations where they are required to do more than merely reproduce the materials from the textbook or elsewhere.

8. The item writer should write essay item in such a way that the task is very clearly defined for the examinees. Here, the aim should be that each examinee's score must reflect how well he can do on the given task and not that how well he can figure out what the task is.

9. When the essay item relates to a controversial issue, it should be evaluated in terms of the presentation of evidence for a position rather than only for taking that position. For example, the essay item such as explain how the idea of live-in relationship fits with Indian culture, should be evaluated in terms of presentation of evidences for a position taken by the examinee and not for the position taken only.

Once the item writer has written the good essay items, he has to do several things that assure that the test will function in the way he wants to do. Following are some of the guidelines that help the item writer prepare the essay test in such a way that will help eliminate some common but unwanted effect on test scores.

1. In preparing an essay test, the test constructor or the item writer must be sure that there are not too many or too lengthy questions for the examinees to answer in the given time. An essay test should not act as an exercise in writing speed. The examinees must get adequate time to think and compose the answers. The more complex questions demand longer time to answer the essay items.

2. In most situations, all examinees should be required to answer the same questions or the same set of questions. Allowing the examinees a choice for answering the questions reduces the common basis for differentiating among them. In fact, comparing the quality of one examinee's answer to questions/item 1 with other examinee's answer to question/item 1 introduces an additional source of subjectivity and measurement error into the process of testing and assessment.

3. As far as possible, the test constructor should have a range of complexity and difficulty among several essay questions or items. Essay tests are designed to make differentiation among examinees in the degree of mastery over the content area. If all questions/items are very difficult, some of the less proficient examinees may remain unable to produce satisfacory answers to the questions. On the other hand, if most questions/items are easy, the test taker will not be able to measure the full capacity of the more proficient examinees. Therefore, a range of difficulty should be provided so that the test may have questions appropriate for all examinees.

4. Test constructor should provide a set of general directions for the entire test rather than providing simple and straightforward directions like, "Answer the following questions". A good set of direction should emphasize upon the following major principles:

(i) The form in which answers are to be provided must be clearly specified.

(ii) A general plan the examinee should take in the class must be outlined.

(iii) The general criteria for making evaluation should also be clearly stated.

(iv) The total time for answering all questions should be clearly written.

(v) The score point to be given for each question or portion of a question should clearly be spelled out.

So far as the scoring of the essay items is concerned, it is a difficult task. In scoring th responses of the essay items, the emphasis should not be placed on factual information becaus for such information objective items are more suitable than essay items. Here greater emphasi should be given to presentation of thoughts and ideas in a coherent manner, critical evaluation and the final judgements of the examinees. There are two general methods of scoring the responses of essay items: the sorting method and the point-score method.

In the *sorting method,* the answer sheets of the examinees are first sorted into different groups according to the fairness of the answers. For example, one group may be designated as the best- answer group, another the better-answer group, and still another as the poor-answer group. There is no specific rule which limits the number of sorting groups. It is dependent upon the number of examinees and the number of discriminations that can legitimately be made among the different answers of the essay items. After this, the scorer assigns the weightage to be given to each group and accordingly, he gives weightage to each answer and finally adds these weights together to constitute a total score for each paper. The advantage of the sorting method is that scoring is done quickly and chances for erratic marking are reduced considerably.

The *point-score method* is well suited to score answers of short-answer essay items. In this method, a grading key consisting of the correct answers and the points to be assigned is prepared beforehand by the scorer. The scorer has only to compare the answers of the examinees with the answers or responses of the grading-key and assign points. Subsequently, all these points are added together to constitute a total score for all the answers in a given paper. The demerit of the point-score method is that it cannot be applied to the extended or long-answer essay as it is difficult to prepare such a grading-key.

In addition to these two methods, some guidelines must be kept in view because these will ease the scoring of essay items. Reality is that a good evaluation of an essay test requires not only a good set of items but also sound and consistent judgement about the quality of answers provided by the examinees. As we know, variations in scoring from one question/item to another question/item and from examinee to examinee contributes to the error in the measurement. Therefore, there are some general guidelines, which if followed, will reduce the probability of committing such errors. These guidelines are as under:

1. The examiner should decide what qualities should be kept in view in judging the adequacy of an answer to each question. If the examiner wants that more than one quality is to be judged, each one should be evaluated separately. For example, if the examinees are to be assessed for grammar and spelling, they should be evaluated separately from when they are to be evaluated for the organisation of facts as well as their knowledge. Of course, the examinees should get information about this in the direction of the test.
2. The examiner or the evaluator should read all answers to one question before going on to the next one. This will serve two purposes. *First,* this will help the evaluator in maintaining a more uniform set of standards across all papers. *Second,* the evaluator is less likely to be influenced in judging the quality of the answer to one item/question by making a comparison to how well the same examinee has answered the other questions.
3. A *model answer* or *Answer guide* should be prepared. Such model answer must contain what points should be covered to receive full marks or credit.
4. The paper should be graded as nearly anonymously as possible. If the evaluator does not know or knows less about who wrote an answer, he can more objectively score the answer.
5. When one question for all examinees have been scored, the papers should be shuffled before starting the process of scoring of the next question. An average examinee who has the misfortune of having his paper scored right after a brilliant examinee, is likely to suffer by the comparison. When the papers are shuffled, the order in which papers appear changes from question to question.

6. As far as possible, the evaluator should write comments and correct errors in the answers provided by the examinees. The rationale behind this principle is that when the examinees get prompt specific information about their strength and weekness, this serves as a motivational device for them.

Objective Item or 'New-type' Item

An objective or new-type item is one wherein there is only one fixed correct answer, which either the examinee gives on his own or he is required to select from among a given few. All objective items can be divided into two broad categories: the *supply type* and the *selection type.* A supply type of item is one wherein the examinee is to write down the correct answer on his own. A selection type of item is one wherein the examinee is to select or identify the correct answer from a few given. Nunnally (1970, 119) refers to such items as *identification items.*

The supply type of item is divided into two main categories: *the unstructured short-answer item* and the *completion item*[*] or *fill-in item.* An unstructured short-answer item is one which is given in a question form and where the examinee supplies the correct answer in terms of a word, number or phrase. A completion item or fill-in item is one which is presented in the form of an incomplete statement. The examinee is supposed to complete the statement by supplying the missing word, number of phrase, etc. Unstructured short-answer items and completion items differ only in form because when an item is written in the question form, it constitutes an unstructured short-answer item but when the same statement is presented in the form of an incomplete statement, it constitutes a completion form of item. The following examples illustrate the two types of items:

1. Who was the first president of the Indian Republic? (Dr Rajendra Prasad) (An unstructured short-answer item)
2. The first president of the Indian Republic was Dr Rajendra Prasad. (A completion or fill-in item)

Of these two types of items, the completion item is frequently used in the field of mathematics, chemistry, physics, and biology where isolated bits of information are to be recalled. Completion items are also used in vocabulary tests and where the association between dates and events is wanted. Completion items have some advantages. They require the examinees to recall their past associations rather than to merely recognize them. It is, therefore, essential that the examinee must have a thorough knowledge of the subject matter. In completion items, the guessing of the correct answer is reduced to a minimum as there are no options or answers given. Completion items have also some disadvantages. Organizational ability requiring complex understanding and applications of scientific and theoretical principles in which synthesis, analysis and evaluation skills are required, is not ordinarily measured by these items. If adequate care has not been taken in writing the completion item, its scoring may not be objective. Sometimes, it happens that the correct answer is not acceptable to many experts because of the inherent defects in the structure of such items. In such a situation, the scoring tends to be subjective. Besides, scoring may get delayed for considerable time. For scoring completion items, a scoring key is prepared in which all the correct answers are listed. The completion items are scored by matching the examinee's responses with the answers given in the key. Usually, for each correct answer a score of +1 and for each incorrect answer a score of 0 is given. The total score is obtained by adding the individual scores. Generally, in completion items no penalty is given for an incorrect answer as chance plays an insignificant role in answering such items.

---

[*] A *completion item* is not as objective as other forms of objective items such as true-false item, multiple-choice items, etc.

In writing the completion items, the item writer must keep in mind the following points:

1. The item must not be ambiguous. Given here are two examples of how the same item can be worded in an ambiguous form and a clear-cut form.

   Ambiguous form: Behaviourism was founded in ____________ .

   The above item has two possible answers: '1913' and 'America', both of which are correct.

   Clear-cut form: Behaviourism was founded in the year _________ .

   or

   Behaviourism was founded in the country known as _________ .

2. Clues which give a direct hint to the correct answer must be avoided. For example:

   A person who believes in God is known as a ____________ and a person who does not believe in God is known as an __________ .

   Obviously, in the first blank "theist" and in the second blank its opposite, "atheist" will be filled in. In such items when the examinee knows the answer of the first blank, he can without much reasoning, answer the second blank because it is the opposite of the answer of the first blank.

3. A large number of blanks should not be used in completion items because in that case originality (and not the real omissions) would be the only source of response.

   For example:

   Personality is the dynamic organization ________ that ________ his _________ .

In the above completion item the examinee is required to give the definition of personality, which in the words of Allport is "the dynamic organization within the individual of those psychophysical systems that determine his characteristic behaviour and thought". Therefore, the above item may be completed as "Personality is the dynamic organization within the individual of those psychophysical systems that determine his characteristic behaviour and thought". In this way, the original idea may not be conveyed if there are too many blanks.

As we have discussed earlier, another important category of objective items is the selection type of item where the examinee is required to select the correct answer from among a few given answers. Selection-type items, as used in psychological, sociological and educational research, are of many varieties. Some of the important varieties are two-alternative items, multiple-choice items and matching items.

1. Two-alternative Item

In this type of item, only two answers are provided from which the examinee is required to select he one which he thinks to be correct. The two-alternative item, also called dichotomous format em, is further divided into four subcategories.

(a) Yes-No: This is very common. Here items are expressed in the form of a question. For example:

1. Do you suffer from insomnia? Yes/No

2. Do you feel shy in talking with the members of the opposite sex? Yes/No

(b) True-False: True-false form of objective item is expressed in the form of a declarative statement, which is either entirely true or entirely false. For example:

1. Wilhelm Wundt is regarded as the father of experimental psychology. True/False

2. Mahatma Gandhi is popularly known as 'Rashtrapita'. True/False

3. Adolf Adler was the father of psychoanalysis. True/False

(c) *Right-Wrong*: Right-wrong form of objective item is expressed in terms of simple sentences, which are to be marked as right or wrong, depending upon the correctness or incorrectness of the language of the sentence. For example:

1. He needs not go to cinema. Right/Wrong
   (The sentence is grammatically wrong.)
2. Mohan said that the sun rose in the east. Right/Wrong
   (The sentence is grammatically wrong.)
3. He said that honesty is the best policy. Right/Wrong

(d) *The Correction Form*: In this form, the examinee is required to correct the item by providing the right answer. Usually, the wrong part of the item is underlined or written in colour or in italics, and the examinee is required to substitute it by a correct answer. For example:

1. The first psychological laboratory was established by Wilhelm Wundt in *1869*. (1879)
2. *January* 26 is celebrated as Independence Day in India. (August 15)

Strictly speaking, the correction form of the two-alternative item is neither the selection type nor the supply type. Rather, it is a combination of the characteristics of both because the examinee supplies the answer on the basis of selection of responses evoked by the wrong answer inserted in the item itself.

The two-alternative form of the objective item is mostly suited to personality tests and tests where noninterpretative *information* like dates, technical terms, and vocabulary is required. The biggest limitation of the two-alternative form is that the probability of guessing the correct answer is 50%. Thus in this form of item, the examinee can select a correct answer on the basis of guesswork for which 50% probability exists. That is the reason why such items are not preferred in the case of intelligence tests and aptitude tests. However, the advantage of the two-alternative item is that it is easier to construct and takes less time in scoring and administration.

There are some general suggestions and guidelines that must be kept in mind at the time of writing two-alternative items.

(a) The statement must be simple and straightforward and not complex and indirect. For example:

Poor form: The experimental study of human behaviour in the laboratory started about hundred years ago.

Improved form: The first experimental study of human behaviour in the laboratory started in 1879.

(b) The statement must be either entirely true or entirely false so that only one interpretation is possible. For example:

Poor form: The capital of India is Delhi. True/False

Improved from: The capital of India is New Delhi. True/False

(c) Double negatives must be avoided because they tend to confuse the examinees. For example:

Poor form: A driver can have no accurate idea of the speed of the vehicle if there is no speedometer in it. True/Flase

Improved form: In the absence of a speedometer, the driver is unable to get an accurate idea of the speed of the vehicle. True/False

(d) The exact authority should be mentioned if the statement of the item is controversial. For example:

Poor form: Intelligence refers to the extent to which a person is able to learn or profit by past experience. True/False

Improved form: According to Ebbinghaus, intelligence is the ability to learn or to profit by past experience. True/False

The first statement, which reflects the viewpoint of Ebbinghaus without referring to his name, is an example of the poor form of the true-false item because the examinee may have to justify his statement. But when the name of the authority is mentioned (as in the second statement), the statement is improved because now there remains little doubt about the validity of the statement.

(e) Specific determiners which are likely to help the examinees in finding out the correct answer must be avoided. Statements which incorporate such words as *never, nothing, always, all* and *none* are likely to be more false than true. When such a word is used in a false statement, it provides additional clues to the correct answer and thus acts as a specific determiner.

Likewise, statements containing *often* and *sometimes* are likely to be more true than false. If these are used in true statements only, they provide additional clues to the correct answer. Hence, such specific determiners must be avoided or a proper balance of each of these two types of specific determiners in true and false statements must be made.

(f) There must be a fixed sequence of the true statement or the false statement when such items are finally arranged in the test. For example, True, True, True, False, True, True, True, False and so on will be an unwise sequence of items but True, False, False, True, True, False, True, False, True, True is an example of a wise sequence of items.

(g) If a statement has been taken from some book or some other source, it should be paraphrased and not reproduced verbatim. The advantage of this is that it will discourage any profit by rote memory. If the subject has read the book from which the statement has been taken verbatim, he will naturally select the correct answer. But when the language has been changed, the examinee will not be able to profit from his rote memory. He can select the correct answer only when he has understood the fundamental principles involved therein.

Multiple-choice Item

is the most popular, common, flexible and effective of all objective items. In each item of the ultiple-choice, there is one stem in the form of a question or incomplete statement, which is nerally followed by a set of three to five answers or options. This is also known as polytomous polychotomous format item. The following examples illustrate a multiple-choice item:

1. *The theory of psychoanalysis was advanced by*
   (a) E L Thorndike (b) I P Pavlov (c) J B Watson (d) S Freud
2. Which of the following is a projective measure of personality?
   (a) Guilford–Zimmerman Temperament Survey (b) Cattell's 16 PF Test (c) Rorschach Test (d) Bernreuter Personality Test

The stem, also known as the premise, has two primary functions: it presents the problem and s the stage for giving answers by providing an appropriate frame of reference. The set of ers or alternatives or options provides the field for operation of that frame of reference which eated by the stem. Although there is no rule for the exact number of alternatives to be ded in the set of options, usually four to five alternatives are preferred. In order that all atives appear equally attractive, they must be of about the same length and complexity.

Among the given alternatives, either one is correct or at least the best among all those given is correct. The examinee is required to indicate that corrrect or best answer. The correct response or the best answer is called the *keyed answer* and the remaining alternatives are known as *distractors or foils*. The aim of foils or distractors is not to confuse the examinee but to provide a very plausible alternative, which he may choose in preference to the keyed alternative if he does not know the correct answer. It is, therefore, essential that distractors must not produce any ambiguity. As a matter of fact, multiple-choice questions are difficult because the distractors and the keyed responses are very similar. The greater the similarity, the more difficult is the item. Multiple-choice items are used in measuring instructional objectives, which include knowledge, vocabulary, analysis, synthesis, application of principles, evaluation, cause-and-effect association, ability to identify error, difference or ability to interpret data. Such items have some marked advantages: they can be used to measure any educational and instructional ability (except the ability to organize the materials) of the examinees of any class or grade. As compared to other objective items, they can be scored rapidly, accurately, and objectively by even those who have little training in the testing field.

Another advantage of multiple-choice item over two- alternative item is that the respondents (or testees) cannot receive credits for simply knowing that the statement is incorrect, rather they must also know what is correct. Look at the difference in the following two items:

1. Mrs Indira Gandhi was assassinated in 1986. True/False
2. Mrs Indira Gandhi was assassinated in
   (a) 1984 (b) 1986 (c) 1994 (d) 1985

In true-false version, the respondents will receive credit if they mark it false even if they think that Mrs Indira Gandhi was assassinated in some year other than 1984. In fact, if a respondent thought that she was assassinated in 1986, it would be marked false and he will receive a credit for that response. Since marking the statement false does not show that the respondent knows what is correct, the resulting score tends to be an inadequate measure of what the respondent knows. But this problem is solved with multiple-choice items because the respondent must select the correct answer to receive the credit.

Another advantage of multiple-choice item is that it is relatively free from response sets. In other words, respondents don't favour a particular alternative when they don't know the correct answer. Still another related advantage is that using a number of plausible alternatives makes the results amenable to some kind of diagnosis. The kind of incorrect alternatives selected by the respondents provides clues regarding the factual terms and misunderstandings that need to be corrected.

Another most commonly cited advantage of multiple-choice item over true-false item is the higher degree of reliability per item. Since the number of alternative is increased from two to three, four or five, the opportunity to guess the correct answer is reduced and therefore, reliability is correspondingly enhanced. In fact, the effect of increasing the number of alternative for each item is similar to that of increasing the length of the test for enhancing the reliability.

Multiple-choice items have a few inherent limitations, too. Such items should not be used to measure the examinee's ability to organize materials or to examine his ability to express himself or for measuring problem-solving skills in mathematics and science. Multiple-choice items are difficult to write, and demand a thorough knowledge of the subject matter, and a detailed knowledge of the art of item writing. This means that multiple-choice items can be written only by trained persons. Such items, as compared to the two-alternative items, take longer time to be answered by the examinee. Another disadvantage of multiple-choice item is related to the finding of a sufficient number of incorrect but plausible distractors. This problem is very serious when the respondents are at primary level, where their vocabulary and knowledge in the area are limited.

An item writer must keep in view the following suggestions while writing multiple-choice items:

(a) The keyed response should be neither shorter nor longer than the other alternatives because in either case, the examinee is likely to get some additional clues to the keyed response.

(b) The keyed response should be randomly varied in its position. Often, it has been reported that the item writer, unconsciously or habitually, places the keyed response of each item either at the top or at the bottom of the set of alternatives. If this is a regular pattern, the examinee is likely to pick up the keyed answer and will respond without much reasoning. Hence, the order of the keyed response should be randomly changed in each item.

(c) The stem of the item must be written in a clear language and it must present a problem, the solution to which is to be found from among the given alternatives. For example:

Poor form: Learning

(a) is change in attitude
(b) is change in retention
(c) is change in behaviour
(d) is change in belief

Improved form: Learning ordinarily refers to

(a) change in attitude
(b) change in retention
(c) change in behaviour
(d) change in belief

In the above two examples, the second one is definitely superior to the first one. In the first example, the stem does not set any problem but gives only a general heading or topic, which is followed by one keyed answer and three distractors. But in the second example, the stem clearly states a problem the solution to which is answer 'c'.

(d) Phonological similarity or clang associations between the stem and the keyed answer must be avoided. Whenever this occurs, the examinee is likely to pick up this association as a possible clue to the keyed answer. For example:

Poor form: Verbal intelligence test measures intelligence on the basis of

(a) Manipulative responses (b) Verbal responses
(c) Oral responses (d) Pictorial responses

Improved form: Verbal intelligence test measures intelligence on the basis of

(a) Manipulative responses (b) Written responses
(c) Oral responses (d) Pictorial responses

The first example is inferior to the second example because there is a clang association between the stem (verbal intelligence test) and the keyed response (verbal responses). When 'written responses' is substituted for 'verbal responses', the clang association is over and the item is improved.

(e) Alternatives must ensure that there is only one correct or best answer. If there is more than one correct or best answer, there will be confusion. For example:

Poor form: In two sets of ranked data, the appropriate measure of correlation is

(a) Pearson *r* (b) Biserial *r*
(c) Kendall *tau* (d) Spearman *rho*

Improved form: In two sets of ranked data, the appropriate measure of correlation is

(a) Pearson *r* (b) Biserial *r*
(c) Point-biserial *r* (d) Spearman *rho*

In the above examples, the first example is inferior to the second example because both (c) and (d) are the correct answers but in the second example, only (d) is the correct answer and hence it is an improved form of the multiple-choice item.

(f) Negative should be used sparingly in an item. Multiple-choice items with words like *Not* or *Except* in the stem, generally increase the reading difficulty of the item and require the examinees to perform more difficult reasoning. When a negative word is included in the stem, it should be underlined or capitalized for attracting the attention of the examinee. Negative items provide little information about exactly what the examinee knows. The use of negative with words like *Not* or *Except* at times is justified because it is important to test whether the examinees correctly know the exception to a general rule or can detect errors.

(g) The item writer must be sure that wrong answers are plausible, that is persuasive but deceptive. As we know, one advantage of multiple-choice items is their capacity to select the correct answer from among three, four or five alternatives and thus, they automatically reduce the chance for guessing the correct answer. This purpose can be served in a better way only when the wrong alternatives of multiple-choice items be attractive and plausible to the examinees who lack knowledge that the item is intended to assess. Therefore, incorrect answers must be logically consistent with the stem and should represent common errors done by the examinees.

(h) The item writer should use the option 'none of these' or 'none of the above' only when the key answer can be taken unequivocally as correct or incorrect. In tests like spelling tests, mathematics test and study skill test an absolutely correct answer can be provided. But in other tests, where the examinee is expected to choose the best answer and where the keyed answer is the best answer, but not necessarily the absolutely correct answer, the use of these two options, that is, *none of these* or *none of the above* is considered not much appropriate. Moreover, the option 'none of the above' should be used when the stem is stated as a question rather than a sentence to be completed. For Example:

Poor form: Of the following, the one that is *not* included in the parts of human brain is:

(a) Chromosomes (b) Plasma
(c) Platelet (d) None of the above

Here the use of the option 'none of the above' is *not* only grammatically inconsistent with the stem but it is logically inconsistent with the stem too. Moreover, the double negative effect of the option and the 'not' in the stem, makes comprehension of the meaning difficult. A better option here would be the use of 'none of the above'. For Example:

Better form: Which is not the part of human brain?

(a) Chromosomes (b) Plasma
(c) Platelet (d) None of the above

The item writer should avoid the use of option 'all of the above' in the multiple-choice test. This type of response option is often used either when only three options can easily be generated and a fourth option is needed, or when there is an easily obtained list of traits or characteristics of something and it is very difficult to come up with three incorrect distractors. In the latter situation, it becomes easy to list three of the characteristics along with the response option of 'all of the above' statement.

So far as the scoring of multiple-choice items is concerned, it is very easy. A separate scoring key is prepared and after matching each response of the scoring key, points are awarded and subsequently these are added to give the total score. Scoring of multiple-choice items can also be done through a punched-out key.

3. Matching Item

In a matching item, there are two columns—right and left. The items on the left column are to be paired with the items on the right column. Items on the left constitute a set of related stems called *premises* and items on the right constitute *options* or *responses* of the item. For example:

In the left-hand column the names of some statistical techniques have been given, and in the right-hand column the names of the persons who are closely associated with these techniques. Match the names with the statistical techniques:

B	1. Product-moment method	A	Fisher
C	2. Rank difference method	B	Pearson
D	3. Coefficient of concordance	C	Spearman
A	4. F ratio	D	Kendall

It is obvious from the above example that the above type of items is quite similar to multiple-choice items. Both the types of items have a list of responses or options. Ordinarily, dates, events, names, technical terms, diagrams, chemical names, etc., are used as stems in matching items and more or less a similar variety is used as the responses of the item. The examinee may be asked to match events with dates, names with events, books with authors' names, terms with definitions, and so on. Thus matching items measure an examinee's ability to associate one thing with another, and hence are effective in measuring achievement where associative learning is involved.

Matching items have, however, some inherent weaknesses. *First,* it is not possible to measure real understanding of concepts through matching items, partly because they simply measure the factual association and partly because they do not measure the organizational ability of the examinee. *Second,* if the examinee knows the correct answers of all pairs except one, he may, through the process of elimination, correct the last pair. *Third,* matching items are not suited when the purpose is to measure different types of information because the construction of such items assumes that the set of stems as well as the set of alternatives would be homogeneous. One different pair among the homogeneous pairs is likely to provide an extra clue to the correct answer and, hence, it is not recommended. *Lastly,* the examinee takes longer to answer matching items as compared to two-alternative items and multiple-choice items.

An item writer must keep in view the following suggestions and guidelines while writing matching items:

(a) The number of alternative answers (or response options) must be greater than the number of stems. This is done in order to reduce the probability of correcting a pair (usually the last pair) through the process of elimination. For instance, when the number of stems is four and the number of alternative answers is also four, and if the examinee knows the correct answers to any three then he will automatically, through the process of elimination, correct the last pair. But when the number of alternative answers is increased from four to six, this probability is automatically reduced.

(b) The item must contain clear instructions or directions regarding the basis of matching, that is, the instructions must state what should be the basis of matching. For example:

Given overleaf are the names of some books and their authors. Match the names of the books with the names of their authors. Matching should be done by putting the appropriate letter in the space provided before each number.

Matching	Authors	Books
B	1. Freud	A. My Experiments with Truth
F	2. Mao Tse-Tung	B. Interpretation of Dreams
A	3. Mahatma Gandhi	C. India Divided
C	4. Rajendra Prasad	D. The Merchant of Venice
D	5. Shakespeare	E. Friends, not Masters
		F. On Contradiction
		G. The Rape of Bangladesh

In the above example there is a clear instruction for matching the names of the books with those of the authors and hence, an examinee will not falter in the process of matching.

(c) The set of stems and the set of response options must be worded in a similar fashion. Placing a different stem among homogeneous stems or a different response option among homogeneous response options would provide a clue. For example:

Poor form: Match the items given in column II with those in column I. Indicate your answer by putting the letter of the correct answer in the space provided before each number of column I:

Matching	*Column I*	*Column II*
C	1. 1821	A. Liberation of Goa
D	2. 1857	B. 1526
A	3. 1961	C. Death of Napoleon
B	4. First battle of Panipat	D. First War of Indian Independence
F	5. 1688	E. Birth of Swami Vivekananda
		F. Bloodless Revolution in England

Improved form: Match the items given in column II with column I. Indicate your answer by putting the appropriate letter in the space provided before each number of column I:

Matching	*Column I*	*Column II*
C	1. 1821	A. Liberation of Goa
D	2. 1857	B. First battle of Panipat
A	3. 1961	C. Death of Napoleon
B	4. 1526	D. First War of Indian Independence
F	5. 1688	E. Birth of Swami Vivekananda
		F. Bloodless Revolution in England

The first example is inferior to the second example because in the first example, column I and II are not homogeneous. In column I are listed the years with an exception in the fourth row (First battle of Panipat) and in column II are listed the names of events with an exception in the second row (1526). Because the name of the event written in column I and the year written in column II are easily identified or distinguished, these two may be paired without real knowledge. Hence, its improved form will be to write the year in column I and the name of the event in column II as has been done in the second example.

(d) The nature of the stems and the number of response options must be neither too short nor too long. This suggestion has two advantages. First, if the matching sets are too long, they may not be homogeneous and the time taken to match them may be longer. Secondly, if the matching sets are too short, the chances for guessing the correct pair are increased, thus lowering the validity of the matching item.

Besides these two general types of formats of items, there are other specific types of formats which are popularly used in the construction of psychological tests. They are:

(i) The Likert format
(ii) The Category format
(iii) Checklists and Q-sorts

The Likert format is a very popular format used in attitude and personality scales. In this format, the test takers or respondents are asked to indicate the degree of agreement with respect to certain attitude objects. This format has been called as the *Likert format* because it was used as a part of Likert's (1932) method of attitude scale construction. In this format, instead of asking for a Yes/No reply, the respondents are given a set of statements which are generally answered in terms of five alternatives: *Strongly agree, Agree, Neutral, disagree* and *Strongly disagree*. These options are scored by giving weights like 5, 4, 3, 2 and 1 respectively for positive items and 1, 2, 3, 4, and 5 for negative items respectively. In some applications, six options are used for preventing the respondent from being neutral. In this case then, the response options will be *Strongly disagree, Moderately disagree, Mildly disagree, Mildly agree, Moderately agree* and *Strongly agree*. A few examples of items based on the Likert format are as under:

Statements	*Response options*				
	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
1. Coeducation improves self-confidence of the person					
2. Coeducation affects character development and morality negatively					

Since responses in the Likert format can be subjected to factor analysis, psychometricians developing the test can easily find groups of items that go together. The scales also require that item analysis should be done to assess item discriminability.

The *category format* is another popular item format used in trait/ability/attitude measurement. It is similar to the Likert format but here the number of choices or options are greater than those of Likert. The most popular category format is the 10-point scale. However, it can have either more or fewer categories. Psychologists have concentrated upon a fact that what should be the optimal number of category points or response points. Some have argued that the optimal number of points should be seven (Symonds 1925) whereas some others have suggested that this optimal number should be three times this number (Champney & Marshall 1939). The reality is that the number of categories required depends upon the fineness of the discrimination respondents are willing to make. If the topic is such about which the respondents are not familiar, they will not be able to make finer discriminations and a scale with few categories will do about as well as a scale with many points. However, when the topic is such that the people are most concerned about it, they will tend to respond best to a greater number of categories. However, for most topics or tasks, a 10-point scale seems to provide enough discrimination (Kaplan & Saccuzzo 2001). Anderson (1991) has also reported that a 10-point scale provides the required discrimation among objects for a wide variety of stimuli. Evidences have also suggested that the number of choices beyond 9 or 10 points may not increase reliability and validity; rather it may pull down reliability because the responses may include an element of randomness when there are so many such alternatives among which the respondents may fail to discriminate (Clark & Watson 1998).

Another approach which is very similar to category scales is the *visual analogue scale*. In this method, the respondent is presented with a 100-millimetre line with two well-defined end points and is asked to place a mark between these two points (Figure 3.1). The scales are scored according to the measured distance from the first end point to the mark.

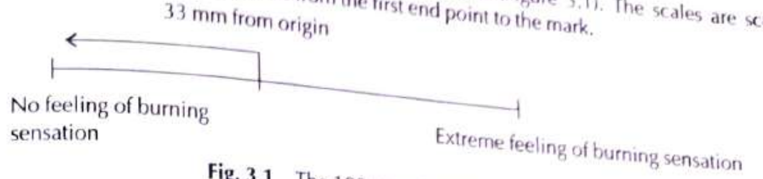
33 mm from origin
No feeling of burning sensation
Extreme feeling of burning sensation

Fig. 3.1 The 100-mm visual analogue scale

In Figure 3.1, the respondent has put a mark at 33 mm from the origin and therefore, his score for feeling of burning sensation is 33. The visual analogue scales are popular for assessing self-rated health. However, they are not used for multi-item scales because their scoring is unnecessarily time consuming (Clark & Watson 1998).

Checklist and Q-sorts are another popular format. In personality assessment, an adjective checklist is very common (Gough, 1960). In an adjective checklist, the respondent receives a long list of adjectives and he has to indicate whether or not each adjective is true in his or her case. Thus in the adjective checklist the respondent is required to either endorse such adjectives or not, allowing only two choices for each item. Such a checklist is used either for describing oneself or someone else. In one study the raters rated the students considered exceptional in originality as 'alert', 'curious', 'imaginative' and 'fair-minded' whereas they rated the students considered low in originality as 'confused', 'polished', 'prejudiced', 'conventional' and 'suggestible'. Q-sort is another similar technique which is used to describe onself or to provide ratings of others (Stephenson, 1953). This is specially used in the study of verbalized attitudes, preferences, self-description, etc. In this technique, the respondents are given a set of statements with the instruction to sort them into nine piles. One basic characteristic of Q-sort* is that the frequency of items place in each category generally and usually looks like a bell-shaped curve.

In this way, there are several types of formats of items and depending upon the purpose, any one is chosen by the researcher. Of these various formats, multiple-choice, Q-sort and Likert formats are most popular in recent times. The other formats are less popular. Today checklists have almost gone out of use because they are more prone to error.

DIFFERENCE BETWEEN ESSAY-TYPE TESTS AND OBJECTIVE-TYPE TESTS

After having discussed the different types of essay-test items and objective-test items, it is essential that we must learn something about the differences between essay and objective types of tests. Following Ebel (1965), the main points of difference may be enumerated as follows:

1. *In essay tests the examinee writes the answers in his own words whereas in objective-type tests the examinee selects the correct answer from among several given alternative answers.*

In essay tests, the examinee is usually required to write down the answer in his own words. This gives him an opportunity to recall and express his ideas in the manner he likes, relate these ideas with some principles, and organize them into a coherent whole. He also gets an opportunity to give his own judgement. An examinee who can express himself well is, thus, likely to score better than the others. However, a common drawback of essay-type tests is that a good handwriting, fluent and graceful structure of sentences may hide some of the defects of the contents of the answers and influence the examiner or scorer's judgement favourably. On the other hand, any minor error in the structure of the sentence or flaw in spelling or usage can unfavourably influence the scorer's judgement regarding the content of the answer.

In objective tests, the examinee is not provided with the opportunity of free expression of his viewpoints. He is simply required to select the correct answer from among the given alternatives.

---

* The details of Q-sort appear in Chapter 12.

Since objective tests give no opportunity for free expression, and the answers are also provided by the test constructor, some experts are of the view that such tests cannot be a good measure of the true knowledge of the examinee. It is also said that the objective items are mere tests of recognition or in some cases, the tests of rote memory only. But these do not seem to be the correct viewpoints because objective-test items can be written in such a way as to stimulate the originality and critical thinking which will be the basis for choice among the alternative answers. Not only this, objective-type test items, if properly planned and written according to the suggestions given earlier, will not permit correct response simply on the basis of recognition or on the basis of rote memory or on the basis of any verbal associations.

2. *Thinking and writing are important in essay-type tests whereas reading and thinking are important in objective type tests.*

In essay-type tests, the examinee answers the questions in several lines. He critically thinks over the problem posed by the question, arranges the ideas in a sequence and expresses them in writing. Thus in essay-type tests, thinking, particularly critical or reflective thinking, and the subsequent ability to express them in writing dominate. In objective tests, the examinee does not have to write much and in many cases he is simply asked to put a tick mark. However, in order to make the correct choice he is required to read both the stems as well as the alternative answers very carefully and then critically think and decide. Thus in objective-test items, reading and reflective thinking dominate.

3. *It is difficult to score essay tests objectively and accurately whereas objective tests can be easily scored objectively and accurately.*

Essay tests are difficult to evaluate objectively, partly because the answers are not fixed like
e answers to objective items, and partly because of the variability in the scorer's judgement
garding the contents of the answers. This variability occurs because the scorer is freely guided
his individual preferences in scoring the answers. Since the answers are not fixed, judgement
rding their accuracy will also vary. In objective-type tests, whether of the selection or supply
scoring can be done accurately because the answers are fixed. The scoring will also be
tive because when the answers are fixed, there will obviously be complete interpersonal
nent (or objectivity) among the scorers.

In objective-type tests, the quality of the items is dependent upon the skill of the test
tor but in essay-type tests, the quality of the items is dependent upon the scorer's skill.

ing items for an objective-type test is a relatively difficult task. Only a skilled test
r can write good objective items. The quality of the test items is bound to suffer if the
uctor lacks skill in writing items or if he has limited knowledge regarding the subject
ns in essay tests are easy to construct. A test constructor even with a minimum
of writing items can prepare relatively good essay items. Whether or not the item has
esired response is judged by a person who evaluates the contents of the answers. So,
gement regarding the good quality of an essay item in producing the desired result
he content of the answer. It is, therefore, often said that a good essay item is not
n the item writer's skill, rather, it depends upon the reader's own skill in judging
worthwhileness of the content.

e-test items, no matter how well they are constructed, permit and encourage
xaminees, whereas essay-test items, no matter how well they are constructed,
age bluffing by examinees.

e test items, the probability of guessing cannot be fully nullified. The effect of
ion of the actual score obtained on the test. Guessing is most obvious when
is short and the two-alternative objective form is used or when difficult
are included in multiple-choice items or matching items and the length of

the test is short. When blind guessing makes a higher contribution towards the make-up of the total test score, the reliability of the scores is considerably lowered. If all examinees guess on all the items of the test, the reliability of the score would be zero.

A similar type of phenomenon observed in the case of essay tests is bluffing. When examinees find essay items difficult, they often try to reinterpret them in a manner which seems easier to them. Not only this, they concentrate more on the length and form of the answer by adding simple facts as well as by deviating from the original idea in the hope that the reader's or scorer's attention will get diverted from the real and true content of the answer. Like guessing, bluffing tends to lower the reliability of the scores. This further tends to deteriorate the reliability of an essay test. It may be recalled here that the reliability of the essay test is already questionable.

6. *Assignment of numerical scores in essay-test items is entirely in the hands of the scorers whereas assignment of numerical scores in objective-type test items is entirely determined by the scoring key of the manual.*

Objective-test items are scored according to a predetermined scoring key prepared by the test constructor. Usually, for each correct answer a score of +1 and for an unanswered or wrongly answered item no credit or score of zero is given. In some objective items, negative marks are given for wrong answers. In this way, the credit to be given and how much credit is to be given are prefixed. But this is not so in the case of essay-test items. Although the test constructor of essay tests prefixes the maximum score to be given in a paper (which is usually 100 or a multiple of 100) and the maximum score to be given to any particular item, how much score should be given in any particular item and what should be the maximum or minimum score assigned to any examinee depends largely upon the will of the scorer. Suppose the maximum marks of an essay item is 20. A person scoring the answer may perhaps find the entire answer correct except the last paragraph which he finds wrong for which he may subtract a score of 9 as penalty and assign a score of 11. There is no rule for penalty. He may subtract even a score of 2 or he may not subtract at all. Likewise, he may assign a total score of 80 on the whole test or he may assign a score of 60 on the whole test. Thus manipulation and assignment of the numerical scores on essay tests are entirely in the hands of the scorer.

Despite the above-mentioned points of differences between essay tests and objective-type tests, they are not altogether different. The main points of similarity are enumerated as follows:

1. An element of subjectivity is involved in both objective-type as well as essay-type tests. In objective-type tests subjectivity is involved in writing the test items, in selecting particular criterion for validation of the test, etc. In essay-type tests, subjectivity is involved in writing and selecting the items. The most obvious effect of subjectivity in essay-type tests is seen in scoring of the essay items.
2. In both essay-type tests as well as objective-type tests, emphasis is placed upon the objectivity in the interpretaion of the test scores. By objectivity is meant that the score must mean nearly the same to all observers or graders who have assigned it. If this is not so, it means that the scoring lacks objectivity, thus reducing the usefulness of the score. Besides, the score must mean the same thing to the person today and several days later.
3. Any educational achievement, such as the ability to spell English words, proficiency in grammar, performance in history, geography, educational psychology, etc., can be measured through both essay-type tests or objective-type tests. When the intention is to measure critical thinking, originality, and the organizational ability, essay-type tests are preferred. But when the intention is to measure piecemeal knowledge in any subject, objective-type tests are preferred. However, this line of demarcation is fast vanishing now because objective items have been used effectively for measuring achievement representing critical thinking and originality of the examinees. Likewise, essay items,

particularly short-answer essay items, have been successfully used in measuring achievement representing piecemeal knowledge of any subject.

GENERAL GUIDELINES FOR ITEM WRITING

The different types of items used in psychological, sociological and educational researches, and some specific guidelines for writing effective items have already been discussed. Besides these specific guidelines, the item writer must also keep in view some general guidelines, which are essential for writing good items. These are listed as follows.

1. Clarity in writing test items is one of the main requirements for an item to be considered good. Items must not be written as 'verbal puzzles'. They must be able to discriminate between those who are competent and those who are not. This is possible only when the items have been written in a simple and clear language. The items must not be a test of the examinee's ability to understand the language. The item writer should be very cautious, particularly in writing the objective items because each such item provides more or less an isolated bit of knowledge, and there the problem of clarity is more serious. If the objective item is a vague one, it will create difficulty in understanding and the validity of the item will be adversely affected. Vagueness in writing items may be because of several reasons, such as poor thinking and incompetence of the item writer.

2. Nonfunctional words must not be included in the item as they tend to lower the validity of the item. Nonfunctional words refer to those words which make no contribution toward the appropriate and correct choice of a response by the examinees. Such words are often included by the item writer in an attempt to make the correct answer less obvious or to provide a good distractor.

3. The item writer must make sure that irrelevant accuracies unintentionally incorporated in the items, are avoided. Such irrelevant accuracies reflect poor critical ability to think on the part of the item writer. They may also lead the examinees to think that the statement is true.

4. The items must not be too easy or too difficult for the examinees. The level of difficulty of the item should be adaptable to the level of understanding of the examinees. Although it is a fact that exact decision regarding the difficulty value of an item can be taken only after some statistical techniques have been employed, yet an experienced item writer is capable of controlling the difficulty value beforehand and making it adaptable to the examinees. In certain forms of objective-type items, such as multiple-choice items and matching items, it is very easy to increase or decrease the difficulty value of the item. In general, when the response alternatives are made homogeneous, the difficulty value of the item is increased but when the response alternatives are made heterogeneous, except the correct alternative, the examinee is likely to choose the correct answer soon and thus, the level of difficulty is decreased. The item writer must keep in view the characteristics of both the ideal examinees as well as the typical examinees. If he keeps the typical examinees (who are fewer in number) in view and ignores the ideal examinees, the test items are likely to be unreasonably difficult ones.

5. Use of stereotyped words, either in the stem or in the alternative responses, must be avoided because these facilitate rote learners in guessing the correct answer. Moreover, such stereotyped words fail to discriminate between those who really know and understand the subject, and those who do not. Thus, stereotyped words do not provide an adequate and discriminatory measure of index. The most obvious way of getting rid of such words is to paraphrase the words in a different manner so that those who really know the answer can pick up the meaning.

6. Irrelevant clues must be avoided. These are sometimes provided in several forms such as clang association, verbal association, length of the answer, keeping a different foil among homogeneous foils, giving the same order of the correct answer, etc. In general, such clues tend to decrease the difficulty level of the item because they provide an easy route to the correct answer. The common observation is that examinees who do not know the correct answer, choose any of these irrelevant clues and answer on that basis. The item writer must, therefore, take special care to avoid such irrelevant clues. Specific determiners like *never, always, all, none* must also be avoided because they are also irrelevant clues to the correct answer, especially in the two-alternative items.

7. Interlocking items must be avoided. Interlocking items, also known as *interdependent items*, are items that can be answered only by referring to other items. In other words, when responding correctly to an item is dependent upon the correct response of any other item, the item constitutes an example of an interlocking or interdependent item. For example:

1. Sociometry is a technique used to study the affect structure of groups. True/False
2. It is a kind of projective technique. True/False
3. It was developed by Moreno, et al. True/False

The above examples illustrate interlocking items. Answers to items 2 and 3 can only be given when the examinee knows the correct answer to item 1. Such items should be avoided because they do not provide an equal chance for all the examinees to answer the item.

Recently, Devellis (1991) has provided some encouraging and stimulating guidelines which are worth considering. Six of these guidelines may be presented as under:

(i) Whatever is going to be assessed should be defined clearly. For this, test developers should use substantive theory as guide and should try to make items as specific as possible.

(ii) Test developers should avoid exceptionally long items, which are rarely considered good.

(iii) The level of reading difficulty should be appropriate for those who will take the test.

(iv) Test developers should develop a pool of items by writing three or four items for each one that will finally be used in the scale. They should avoid redundant items. Although theoretically all items are randomly chosen from a universe of item contents, practically, however, care in selecting and developing items is valuable.

(v) 'Double-barreled' items should be avoided. Such items convey two or more ideas at the same time.

(vi) Negative and positive items should be mixed. This is needed to check up 'acquiescence response set' which indicates that the respondents develop a tendency to agree with most items. To check or avoid such response bias, the test developers should write the same items in the negative direction. For example, in writing items about depression, the positive item would be: "I feel sad" whereas the negative item would be: "I feel hopeful and delightful."

Apart from these broad suggestions, test developers should keep in view that with lapse of time, tests lose their half-life and about 12% of items become less reliable over this time (Chan, Drasgow & Sawin 1999). In general, those items tend to lose reliability over time which are focused on the meaning of a concept.

GENERAL METHODS OF SCORING OBJECTIVE-TEST ITEMS

Scoring of the test items is one of the important aspects of objective-type tests. There are several methods for scoring the responses. The following factors must be kept in view while selecting a method of scoring:

1. It must not take much time.
2. It must not be such that only a highly trained person can adapt it.

3. It must ensure objectivity. Objectivity, here, means interpersonal agreement among the different scorers. If different persons are scoring the same test, the method of scoring must be such that all of them arrive at the identical conclusion.
4. The method of scoring should produce maximum accuracy.

The following are the important methods of scoring objective-test items.

Overlay-key Method

In this method, a cut-out key is prepared, that is, a window which may display the answer to each item on each page is prepared. The correct answers are written just below the windows. Subsequently, the key is overlaid upon each page in such a manner that through the window the answer (to an item) given by the examinee can be seen and compared with the correct answer written below the window. The examinee's total score is the sum of the points given for each correct answer. Such a method can also be used for scoring matching and completion items.

Strip-key Method

In this method, the correct answer to each item is printed vertically on a strip of paper on a cardboard. The answers are printed in such a way that they run parallel to their respective items of the test. After matching the answers (given by the examinee) with the answers printed on the strip, they are marked in an adjacent vertical column on the left of the respective items. The examinee's total score is the sum of the points given for each correct answer.

So far as the scoring of essay-type tests is concerned, no such key-method can be developed. However, short-answer essay items can be scored like objective items if a certain planning like the length of the answer, the type of factual information to be included, the weight or points to be given to each type, model answers, etc., is done beforehand. A similar planning can also be done with the scoring of the extended-answer essay items, but how far it will be successful is difficult to tell.

☑ Review Questions

1. Discuss the meaning of test item. Also discuss the relative advantages and disadvantages of different types of items.
2. Distinguish between essay-type tests and objective-type tests. Cite appropriate examples.
3. Outline the major precautions that should be taken in writing test items.
4. Which of the different methods of scoring objective-test items do you consider best? Give reasons for your answer.
5. Discuss the advantages and limitations of the different types of objective items.

◆

4

ITEM ANALYSIS

CHAPTER PREVIEW

- Meaning and Purpose of Item Analysis
- Power Tests: Item Difficulty
    - Index of Dicrimination
    - Distractor Analysis
- Optional Difficulty Value for a Reliable Test
- Index of Discrimination
- A Simplified Item Analysis Form
- Effectiveness of Distractors or Foils (Distractor Analysis)
- Speed Test
    - Index of Difficulty
    - Index of Discrimination
- Factors Influencing the Index of Difficulty and the Index of Discrimination
- Problems of Item Analysis
- Important Interactions Among Item Characteristics
- The Item Characteristics Curve (ICC) and Item Response Theory

MEANING AND PURPOSE OF ITEM ANALYSIS

After the items have been written, reviewed and carefully edited, they are subjected to a procedure called *item analysis*. Item analysis is a set of procedures, that is applied to know the indices for the truthfulness (or validity) of items. In other words, item analysis is a technique through which those items which are valid and suited to the purpose are selected and the rest are either eliminated or modified to suit the purpose. For the purpose of selecting the most appropriate, valid and discriminating items the individual item performance by a group of examinees is compared to their performance on the whole test. In brief, it can be said that item analysis demonstrates how effectively a given test item functions within the total test. As a matter of fact, each item is an element within the test and in this sense, it is a test within a test. The validity of the whole test is dependent upon the validity of the individual items. Item analysis is a set of procedures that provides us with the estimate of validity of each item. The main objectives of item analysis are enumerated below.

1. Item analysis indicates which items are difficult, easy, moderately difficult or moderately easy. In other words, it provides an index of the difficulty value to each item.
2. It also provides indices of the ability of the item to discriminate between inferior and superior. In other words, item analysis indicates the discrimination value of each item. This is known as item validity.
3. It indicates the effectiveness of the distractors in multiple-choice items. Since multiple-choice items are one of the most powerful and flexible objective items and many of the standardized tests utilize this form of item, a thorough item analysis is done to indicate the extent to which the distractors or foils are effective in each item.
4. It sometimes also indicates why a particular item in the test has not functioned effectively and how this might be modified so that its functional significance can be increased.

POWER TESTS

A power test is one where the examinee is allowed sufficient time for answering all items of the test. Thus, emphasis here is upon measurement of the ability (or power) of the examinee and not upon his speed. Ideally, in such a test, items are of different difficulty values and they are arranged in increasing order of difficulty.

The two most common indices which item analysis reveals are the index of difficulty value and the index of discrimination value. Item analysis also aims at knowing the effectiveness of distractors in multiple-choice items. This part of item analysis is known as distractor analysis. There are different methods of calculating the difficulty value and discrimination value of an item. A detailed discussion of these methods is presented below.

ITEM DIFFICULTY

In item analysis, the first step is to find out the difficulty value of the item or the index of difficulty of an item. The difficulty value of an item is defined as the proportion or percentage of the examinees or individuals who answer the item correctly. This proportion or percentage is known as the index of difficulty of an item. If an item is answered correctly by 90% of the examinees, it obviously means that the item is relatively easy and is not discriminating well at the very bottom of the scale of the trait being measured, as it is discriminating only in 90 × 10 combinations or pairs or examinees. Likewise, if an item is answered correctly by only 3% of the examinees, it means that the item is a difficult one and is not discriminating well among the examinees at the higher level of the trait being measured or is discriminating only among superiors and genius examinees. Also, an item which is answered correctly by 100% and 0% examinees has no differentiating significance. The maximum number of possible discriminations between examinees of any item is 50 × 50 = 2500. This occurs when an item is answered correctly by 50% of the examinees and answered wrongly by 50%. But this does not mean that the tests composed of items of only 50% difficulty level are always to be preferred because such tests would simply divide the total examinees into two groups: those above the 50% difficulty level and those below the 50% difficulty level. Such items would not differentiate between examinees in a group above or below the 50% level of difficulty. The proportion passing an item is inversely related to the difficulty of an item. The higher the proportion or percentage getting the item right (i.e., the higher the index of difficulty) the easier the item, and the lower the proportion getting the item right (i.e., the lower the index of difficulty) the more difficult the item. If a test is composed only of items of higher indices of difficulty or only of lower indices of difficulty, it will never be a good measuring instrument. Any test to be called a good measuring instrument must have some items of higher indices of difficulty which should usually be placed at the beginning of the test, some items of moderate indices of difficulty (ranging from 40% through 60%) which should preferably appear after that, and some items of lower indices of difficulty which should appear last of all. At the beginning of the test can be placed those items which are passed by 100%, while at the end, those items can be placed which are passed by no one or zero per cent. Thus the test items must have a normal distribution with respect to their indices of difficulty. The reason why items of moderate difficulty indices (which usually range from 0.40 through 0.60) are preferred is that they signal the maximum variance (or variability among the examinees). Variance is defined as SD^2 and the standard deviation (*SD*) of an item which is scored either as 1 or 0, is equal to $\sqrt{pq}$ where *p* is the mean of proportion passing an item and *q* the proportion failing an item. There is a general consensus among experts that for the sharpest discrimination among examinees, the index of difficulty should be 50% or 0.50. If *p* of an item is 50% or 0.5, *q* is also, then, naturally 50% or 0.5. Therefore, its $SD = \sqrt{0.5 \times 0.5} = \sqrt{0.25} = 0.5$ which, when squared, becomes the variance which is equal to 0.25. Directly, it can also be said that the variance of an item is equal to $p \times q$ in which *p* value is the mean on the dichotomous item and *q* is equal to $1 - p$. Variance is directly related to *p* and it is the largest when *p* is close to 0.5. Thus, the maximum variance that an item can yield is 0.25. Maximum variance indicates maximum number of comparisons or separations among individuals who pass and fail. For example, if an item is passed by 50% in a group of 100 examinees, it means it discriminates between 50 who pass it and 50 who fail it. The total number of comparisons is, therefore, equal to 50 × 50 = 2500. Likewise, if the item is passed by 60%, the maximum comparison that this item is capable of yielding is equal to 60 × 40 = 2400. Thus, the best item in terms of comparison or variability is the item whose index of difficulty is 0.50 or 50%. As the index increases or decreases, the variance of the item gradually decreases, that is, its ability to make comparisons among those who pass and those who fail decreases.

Basically, there are two important methods of determining the difficulty value of an item: the method of judgement and the empirical method. A brief discussion of each of these two is presented below.

Method of Judgement

In this method, the difficulty value of an item is determined on the basis of the judgement of experts. Items are given to a group of experts with the instruction to rank the items in their increasing order of difficulty. Subsequently, the test constructor takes a final decision, keeping in view the commonality of ranks assigned to each item by the different judges or the experts. Sometimes, an estimate regarding the index of difficulty is made by the test constructor himself on the basis of the total time taken by the examinees in its solution. The demerit of both these types of procedure is that they cannot be fully reliable and objective.

Empirical Method

The empirical method, also called the *statistical method*, is the basic and scientific method of determining the index of difficulty of an item. There are two common statistical methods through which the index of the difficulty can be estimated. One method determines the index on the basis of the response of all the examinees whereas another method determines the index on the basis of the responses of only a portion of the examinees. Usually, the correct answer in the test is given a score of 1 and the incorrect answer a score of 0. Such items are known as dichotomous items which are different from multipoint items where different points are given as scores, e.g., a score of 5, 4, 3, 2 and 1 may be given to the response categories *strongly agree, agree, neutral, disagree* and *strongly disagree respectively*. A score of +1 on the ability test indicates that the examinee has passed the item and zero means he has failed the item. On the nonability test (such as on the test of personality, attitude, etc.), a score of +1 indicates the highest score and a score of zero indicates the lowest score on any given item. For a dichotomous item where correct answer is scored 1 and incorrect answer is scored 0, the index of difficulty can be determined by the following formula:

$$p = \frac{R}{N} \tag{4.1}$$

where *p* is the index of difficulty; *R* is the number of examinees who pass the item (or correctly answer); and *N* is the total number of examinees who take the test.

The difficulty or *D* values are important for two primary reasons. *First*, *D* value directly influences the shape of total test scores as well as their standard deviation. If the test consists of items of higher *D* values (i.e., easier items), the distribution of scores would be negatively skewed and the standard deviation would be small whereas if the test consists of items of lower *D* values (i.e., difficult items), the distribution of total scores would be positively skewed and the standard deviation would again be small. Since it is desired that the distribution of the total scores be symmetrical and the items disperse examinees as far as possible, the test should preferably consist of items having *D* values close to 0.5. Such items will procure symmetrical distributions as well as the larger standard deviation. A test which consists of items having *D* values close to 0.5 (that is, in the range of 0.4 to 0.6), is known as a *peaked test*. *Second*, *D* values influence the reliability of the test which is statistically defined as the self-correlation of the test. If the items have extreme *D* values (such as 0.2, 0.1, 0.8, 0.9, etc.), they are not likely to correlate with each other, that is, the inter-item correlation will be very low. As a consequence, the test will have a lower internal consistency reliability coefficient. If the items, on the other hand, have *D* values of middle range (e.g., from 0.4 to 0.6), they will be expected to correlate with each other though not perfectly because only items having very close *D* value of middle range, are expected to correlate perfectly. Naturally, then, items having *D* values of middle range are likely to enhance the internal consistency reliability coefficient of the test. Thus, good items are those which have difficulty values near 0.5 because they correlate moderately with each other. This correlation has another implication. The correlation of an item with total scores on the test is directly related to the correlation of a particular item with other items in the test. Items having high inter-item correlations are expected to correlate with total scores of the test. Thus the *D* values not only affect the reliability coefficient of the test but also the item-total score correlation.

The index of difficulty can also be determined from a certain portion of the group of examinees. If the number of examinees is ranked from highest to lowest on the basis of the total score, the upper 27% and the lower 27% may be separated for the purpose. The upper 27% may be called the upper group and the lower 27% may be called the lower group. Thus, from the total number of the examinees the middle 46% are set aside and only on the basis of 54% the index of difficulty of each item is determined. The index of difficulty determined on the basis of these two extreme groups is almost the same when the entire sample of the examinees is taken into account. The formula for determining the index on the basis of the extreme groups is given below:

$$p = \frac{R_u + R_L}{N_u + N_L} \tag{4.2}$$

where *p* is the index of difficulty; R_u is the number of examinees answering correctly in the upper group; R_L is the number of examinees answering correctly in lower group; N_u is the number of examinees in upper group; and N_L the number of examinees in the lower group.

To illustrate, suppose $N_u = 200$, $N_L = 200$, and $R_u = 150$ and $R_L = 50$ for a particular item in a test then its index of difficulty will be as follows:

$$p = \frac{150 + 50}{200 + 200} = \frac{200}{400} = 0.50$$

There is a still easier way of determining the index of difficulty of an item when two equal extreme groups have been set up. On each item, the number of correct responses in the upper group and the lower group is counted and converted into proportions by dividing them by the *N* of the corresponding group. Thus for each item, the correct proportion in the upper group and correct proportion in the lower group are known. Then, by the simple process of averaging the two proportions, we get *D*. For example, the number of examinees giving correct answers in the upper group is 150 and in the lower group it is 50 only. Converting these two into proportions, we get ${}^{150}\!/_{200} = 0.75$ and ${}^{50}\!/_{200} = 0.25$ respectively. Averaging these two proportions, we get

$$(0.75 + 0.25)/2 = 0.50$$

The extreme-group method of determining the index of difficulty of an item is particularly referred where the index of discrimination of the item is also to be determined on the same xtreme groups. This saves labour and time in item analysis.

It has been observed that sometimes examinees answer correctly on the basis of guesswork. essing not only inflates the actual score but also introduces a measurement error. The bability of guessing in any objective-test items cannot be ruled out completely. In the

two-alternative item, the problem of guessing is more acute but in the multiple-choice item the problem of guessing is automatically minimized. The effect of this guessing is to inflate the actual total score on the test and increase the index of difficulty. The amount of increase in the index of difficulty is inversely related to the number of alternative responses for each item. The larger the alternatives, the poorer is the increase in the index of difficulty due to poor chance success on the basis of guessing. It is, therefore, advised that a correction for chance success or guessing should be introduced and then on the basis of the corrected score, the difficulty value should be determined. In applying correction for guessing, it is assumed that the examinees give wrong answers due to absence of real knowledge and for such examinees all response options are equally attractive. If the index of difficulty is being determined on the basis of the upper group and lower group, it is proper first to correct the proportion of the correct answers of the upper and lower groups and then determine the index of difficulty or discrimination. The formula for correcting the total score for chance success is given below:

$$S = R - \frac{W}{K - 1} \tag{4.3}$$

where *S* refers to the corrected score; *R* refers to the number of correct responses; *W* refers to the number of incorrect responses; and *K* refers to the number of response options or choices used in the item. One assumption of the above formula is that all incorrect responses result from guessing. Examinees are expected to either get an answer correct or omit it. To illustrate the computation from the above formula, suppose an item is correctly answered by 300 examinees out of 400, and the remaining 100 answer it incorrectly, and each item has five response options having one score for each correct answer. Instead of a total score of 300, now the corrected score on that particular item will be $300 - [100/(5-1)] = 275$. Accordingly, *P* from ${}^{300}\!/_{400} = 0.75$ will be reduced to ${}^{275}\!/_{400} = 0.6875 = 0.69$. Likewise, in the example cited above where N_u and L_u were used, the number of R_u was 150 and R_L was 50 on the said item. Since 150 examinees in the upper group answered correctly, it means 50 examinees answered incorrectly and 50 in the lower group answered correctly, which means 150 answered it incorrectly. Further, suppose that the item had six response options. Now, the corrected score in the upper group and lower group will be as follows:

$$S = 150 - [50/(6-1)] = 140.$$

So, the corrected score of R_u on the said item is 140 instead of 150.

$$S = 50 - [150/(6-1)] = 20.$$

The corrected score of R_L is 20 instead of 30 on the same item. Accordingly, *D* which was 0.50 will be reduced as shown below:

$$p = \frac{R_u + R_L}{N_u + N_L} = \frac{140 + 20}{200 + 200} = \frac{160}{400} = 0.40.$$

Thus, the corrected *p* is 0.40 instead of 0.50. Guessing can also be minimized through some nonstatistical methods. The examinees should be given a stern warning against guessing so that they resort to guessing the least when they do not know the correct answer. The distractors of the items should be made attractive to those who do not know the correct answer; and finally, every examinee should be given sufficient time for answering each item of the test.

The purpose of introducing correction for chance success or guessing is manifold. *First*, it is done to discourage guessing among the examinees. *Second*, it is done in order to give an apparent advantage to the examinees who don't guess over those who guess. *Third*, it makes the score meaningful and dependable. *Fourth*, the correction for guessing makes even those items comparable directly which have a different number of distractors.

The correction for guessing formula has been criticised because of its assumptions, which are difficult to meet. Moreover, its assumption also fails to do what it intends to do (that is, to make correction for guessing). As we know, the formula is based upon its two fundamental assumptions as under:

(i) One assumption is that all guesses done by the examinees are blind and each response option has an equal probability of being chosen.

(ii) Another assumption is that every wrong answer results from guessing. In other words, the examinee never responds on the basis of misinformation or partial information.

Let us analyze these assumptions. A close attention to the first assumption reveals that it rejects the probability that an examinee might make a clever and educated guess based on partial information or he might eliminate some options. If this really happens, there is need to bring a change in the correction for guessing formula. The second assumption expresses the possibility that an examinee might make a wrong answer based on misinformation, partial knowledge or misunderstood knowledge. But the reality is that a good test constructor generally provides misleads (wrong but similar to a correct answer) so as to attract examinees with partial or incorrect information.

Besides these, the problem of *undercorrection* and some *technical factors* are also associated with correction for guessing formula. On a multiple-choice test with alternative options, the formula may undercorrect to the extent an examinee can eliminate some options and then guess between the remaining number of options. If even one option is eliminated, guessing is likely to help the examinee's score. On the other hand, some examinees who are not much familiar with knowledge of objective tests, are likely to be intimidated by knowing the fact that the correction for guessing formula is employed and due to this realization, they may keep themselves out of guessing work.

Another important characteristic of the correction for guessing formula is that it fails to change the rank order of the examinees, if no questions/items are omitted. In such situation, there will be a perfect correlation (1.00) between the corrected scores and the raw scores. When there are some omitted responses, the coefficient of correlation becomes less than 1.00 and even in such situation, it remains in 0.90s if there is not much variations in the number of items omitted (Ebel 1979). As we know, reliability and validity are increased by correction for guessing only when there is a positive correlation between raw score and omitted responses (Ebel 1979). Of course, such circumstances normally don't occur because the examinee who answers most questions and works quickly is also more knowledgeable and testwise.

Some experts have suggested that intercorrelations of all items should be computed because it gives an indication regarding the distribution of item difficulties, which, in turn, helps in the process of item selection. Under such circumstances, the index of difficulty of the item is not considered separately as it automatically contributes to inter-item correlation. The best index of item intercorrelation is the phi-coefficient if the items are scored as either + 1 or 0. When items are nearly equal in difficulty, their intercorrelation as measured by the phi-coefficient is high (and therefore, reliability of such a test is high). When items vary in the indices of difficulty, such items don't yield high intercorrelations and this yields a low reliability for the test.

As stated earlier, items having difficulty value close to 0.5 are generally preferred. But there are also some disadvantages associated with items having indexes of difficulty close to 0.5. On the attitude scale, the interest test and the personality test in which items have commonly two-alternative options such as *agree-disagree*, *true-false*, *yes-no* and the like, only the most ambiguous and vague items will yield the difficulty value of 0.5 or near that. A test composed of such items would have inter-item correlation near zero and hence, the reliability of the test would be zero. On the ability test, however, *D* values of 0.5 are the most preferred ones. Likewise, in certain cases the *D* value of 0.5 or near that does not always guarantee a high correlation with total scores. When, for example, the index of difficulty of an item has been determined by averaging the proportions of the correct responses in the upper 27% and the lower 27% of the total examinees, and in the upper 27% the proportion correct is 0.5 and in the lower 27% the proportion correct is also 0.5 (so that the index of difficulty is 0.5), its correlation is zero (cf. Table 4.1, p. 66 in which item 7 has item total correlation of zero because of the reasons stated above.)

OPTIMAL DIFFICULTY VALUE FOR A RELIABLE TEST

Generally, the test constructor is faced with the question: What should the optimal level of difficulty of the items be? The question can be answered by referring to what is meant by optimal level of difficulty. Obviously, most test constructors mean that optimal level of difficulty is one which produces maximum item intercorrelations, and thereby maximum item-total correlation and also a higher reliability of test scores. So, the optimal level has a direct link with the reliability of the test scores. Difficulty levels producing higher reliability are the optimal levels. In the discussion on the index of difficulty, it was pointed out that the best index of difficulty is 0.5 or those which are very close to that. This conclusion should not, however, be taken as a *general rule*. When the number of items in the test is 30 or less, items having indexes of difficulty 0.5 or close to it, yield higher intercorrelations and/or item-total correlation and, therefore, also yield higher reliability. But in the longer tests where items are more than 30, scattered indexes of difficulty are preferred though they do not yield a higher inter-item correlation. In *true-false* items, indexes of difficulty concentrating at 0.5 are not good because this would simply mean that items were so difficult that every examinee had guessed. A test consisting of only such items would have a zero reliability coefficient. Likewise, a personality test composed of agree-disagree items having indexes of difficulty at 0.5, is not a good test because it would mean that items were highly ambiguous. It is, therefore, recommended that 'true-false' items or 'agree-disagree' items should have *P* value above 0.70. Lord (1952) has suggested that the optimal level of difficulty for two-alternative items is 0.85 (uncorrected), for three-alternative items is 0.77, for four-alternative items is 0.74 and for five-alternative items 0.69.

INDEX OF DISCRIMINATION

The determination of the index of discrimination also known as the *item validity index* is another important aspect in item analysis. The index of discrimination has been exhaustibly described by Mosier & McQuitty (1940), Johnson (1951) and Ebel (1965). Since then, it has become the vogue to determine this index in any item-analysis work. Index of discrimination is that ability of the item on the basis of which the discrimination or distinction is made between superiors and inferiors (Blood & Budd 1972). Bean (1953, 153) has defined this index as "the degree to which the single item separates the superior from the inferior individuals in the trait or group of traits being measured." Marshall & Hales (1972, 81) have said, "The discriminatory power or validity (*V*) of the item may be defined as the extent to which success and failure on that item indicate the possession of the trait or achievement being measured". From the point of view of discriminatory power, all test items can be divided into items that are either (a) positively discriminating, (b) negatively discriminating, or (c) nondiscriminating. A positively discriminating item may be defined as one in which the proportion or percentage of correct answers is higher in the upper group. A negatively discriminating item may be defined as one in which proportion or percentage of correct answers is lower in the upper group. Likewise, a nondiscriminating item may be defined as one in which the percentage or proportion of correct answers is equal or approximately equal in both the upper and lower groups. The negatively discriminating items as well as the nondiscriminating items are dropped after item analysis because they do not make a positive contribution to the overall functioning of the test.

There are two common ways of determining the index of discrimination:

(i) By applying the test of significance of the difference between two proportions or percentages

(ii) By applying correlational techniques.

A detailed discussion of these two methods of determining the item validity is given below.

(i) Test of Significance of the Difference between Two Proportions Or Percentages

In this method of determining the validity index of an item, the examinees are divided preferably into two equal groups on the basis of the total score. For this purpose, we can take top and bottom 40% or 50% and compare the proportions of answering it correctly in each of these two extreme groups. But the common observation has been that when the dichotomy is made near the median (say 40% top and bottom or 50% top and bottom), it has little influence upon the discrimination index. Kelley (1939) has demonstrated that when the extreme groups consist of top 27% and bottom 27% in a normal distribution, the ratio of difference between the means of the two groups over the standard error of the difference between these two means (that is, the critical ratio) is at a maximum. Kelley has, therefore, recommended that for item analysis N should be 370 so that in the upper 27% there will be 100 examinees and in the lower 27% there will also be 100 examinees. When the distribution is not normal, that is, flatter than the normal one, the optimum percentage for the extreme groups is slightly higher than 27% and reaches approximately 33% (Cureton 1957). Many experts have been influenced by this finding of Kelley and since then, it has become a fashion to use the upper 27% and the lower 27% for determining the validity index of an item. The additional advantage in this approach is that on the same extreme groups, the index of difficulty can also be calculated and thus, both time and labour can be saved in the process of item analysis.

For the purpose of making two extreme groups, the total scores are arranged from highest to lowest and then the top 27% and the bottom 27% are selected. Thus, the total score is the criterion for marking these two extreme groups. When the proportions of examinees passing in the upper group (R_u) and in lower group (R_L) have been determined, critical ratios can be applied to test the significance of the difference between the proportions of these two groups compared with their own standard errors. If the difference comes to be a significant one, the item is accepted as one which discriminates. The demerit of the critical ratio is that it only indicates whether or not the given item discriminates but it does not tell how well each item discriminates, that is, the amount of discrimination power of each item is not revealed.

Guilford (1954) has recommended the use of the chi-square test as a measure of the index of discrimination when there are equal numbers of examinees in both the extreme groups. He has suggested a special reduced formula of chi-square for this purpose:

$$x^2 = \frac{N(p_u - p_L)^2}{4pq} \qquad (4.4)$$

where N refers to the number of total examinees; p_u and p_L refer to the proportion of examinees passing in the upper group and lower group respectively; P is the arithmetic mean of p_u and p_L and q is equal to 1 – p. However, the use of chi-square as a measure of the index of discrimination has some limitations. The chi-square can be used only with larger samples and with items falling in the middle range of attractiveness for the examinees. Moreover, the chi-square test only reveals whether or not the item is discriminating. It gives no idea of the amount of discrimination of each item.

Marshall & Hales (1972) have suggested a very simple and quick method of determining the index of discrimination. They have called this index the "Net D index of discrimination". They have defined Net D as "an unbiased index of absolute difference in the number of discriminations made between the upper group and the lower group—it is proportional to the net discriminations made by the item between the two groups". Like the methods discussed above, the Net D method also demands the setting up of two extreme groups—one consisting of the upper 27% and the other consisting of the bottom 27% of the examinees. This method, as is clear from the following formula, is directly based upon the difference between the proportion of correct answers of the top 27% and the bottom 27% examinees.

$$V = \frac{R_u}{N_u} - \frac{R_l}{N_l} \qquad (4.5)$$

$$\text{Or,} \quad V = \frac{R_u - R_l}{N_u}, \text{ because } N_u = N_l$$

where V = Net D; R_u and R_l refer to the number of the examinees giving correct answers in the upper group and the lower group respectively; and N_u and N_l are the number of examinees in the upper group and in the lower group respectively.

When the value of V is negative, the item displays negative discrimination and thus, it is dropped from the final form of the test. When the values of V (Net D) of items are above 0.40, they are thought to be discriminating well between superior and inferior. Net D falling between 0.20 and 0.40 are considered to be discriminating poorly and those falling below 0.20 are considered to be the index of very little discrimination or negligible discrimination. To illustrate, suppose the number of correct responses in the upper group is 80 and the number of correct responses in the lower group is 70 on an item; and the total number of examinees in each of these two extreme groups is 100. Now the Net D may be determined as follows;

$$V = \frac{R_u}{N_u} - \frac{R_L}{N_L} = \frac{80}{100} - \frac{70}{100} = \frac{80-70}{100} = \frac{10}{100} = 0.10$$

$$\text{Or,} \quad V = \frac{R_u - R_L}{N_u} = \frac{80-70}{100} = \frac{10}{100} = 0.10$$

The said item has the Net D of 0.10, which indicates that the item has negligible discriminatory power. Such items are ordinarily dropped or suitably modified.

Correlational Techniques

Correlational techniques have also been frequently employed as the measures of the index of item discrimination. In such situations each item is correlated against the internal criterion of the total score, that is, each item is validated against the internal criterion of the total score. This is called *item-total correlation*. Lindquist (1951) calls this the "internal-consistency item discrimination index". When the correlation between the total score and the individual item score is computed as a measure of the discriminative power of the item, it shows how well the item is measuring that function which the test itself is measuring. The validity index exhibits the extent to which a particular item discriminates among examinees who differ sharply in the function measured by the test as a whole. As a matter of fact, item-total score correlation is regarded by most of the experts as the best index of discrimination. In selecting items on the basis of item-total correlations, it is better if at least 75% of correlation is positive and preferably above 0.15. Five common measures of correlation such as Product Moment (or PM) correlation, biserial r, point biserial r, tetrachoric r and phi-coefficient are frequently applied in determining the index of discrimination of an item. When multipoint items are employed, the product-moment correlation is the most appropriate one. When items are in two-alternative responses or somehow reducible to two-alternative responses, that is, when dichotomous items are employed, biserial or point biserial r is the most preferred statistic. When the total score is also dichotomized as the response items, the tetrachoric r or the phi-coefficients are used.

The various correlation techniques described above involve a good deal of avoidable routine work. Flanagan was the first person to provide a short cut to these correlation coefficients, especially the biserial *r*, and thus, avoid the unnecessary routine work involved in its calculation. He prepared an abac table* from which one can directly read the value of biserial coefficient provided one knows the proportion of examinees passing an item in the upper 27% criterion group and the proportion of the examinees passing in lower 27% criterion group. In each extreme group *N* should not be less than 100, which means that the total number of examinees should be 370.

Table 4.1 Item analysis worksheet

1	2	3	4	5
Item No.	*N* = 100 proportion correct in upper 27%	*N* = 100 proportion correct in lower 27%	Index of difficulty	Index of discrimination
1	0.80	0.20	0.50	0.60
2	0.65	0.35	0.50	0.30
3	0.90	0.40	0.65	0.50
4	0.40	0.60	0.50	–0.20*
5	0.65	0.85	0.75	–0.25*
6	0.55	0.25	0.40	0.30
7	0.50	0.50	0.50	0.00*
8	0.70	0.70	0.70	0.00*
9	0.75	0.05	0.40	0.70
10	0.95	0.90	0.92	0.05*

* Items either to be dropped or to be modified

The data in Table 4.1 illustrates the computation of biserial *r* with the Flanagan abac table on the basis of the top 27% and the bottom 27% of the examinees (separated on the basis of the total score distribution).

In Table 4.1, the item analysis of only 10 items of a test has been shown. In column 2, the correct proportion of those passing the items in the upper 27% of the group and in column 3, the correct proportion of those passing the items in the lower 27% of the group have been shown. (Thus the middle 46% of the examinees have been excluded.) In column 4, the index of difficulty of each item has been shown. The index has been found by averaging the proportions of columns 2 and 3. Item 1 has 0.80 and 0.20 as correct proportion in the upper and the lower groups respectively. The average of these proportions is equal to $(0.80 + 0.20)/2 = \frac{1}{2} = 0.50$. In this way, the difficulty value has been determined. The difficulty value in column 4 is for the entire examinees calculated on the basis of 54% (top 27% plus bottom 27%) of the group. The index of discrimination of each item appears in column 5 which exhibits biserial *r* read directly from the Flanagan abac table. In the graph of this abac table, the correct proportions of the upper group are shown on the ordinate and the correct proportions of the lower group on the abscissa. The lines showing these two proportions intersect at a point in the graph. The point of intersection is the desired coefficient of biserial *r* (cf. Guilford 1954, 428). These correlations can also be computed using Equation 4.5.

It is obvious from the above worksheet of item analysis that items 4 and 5 have negative coefficient indicating discrimination and items 7 and 8 have zero coefficients indicating no discrimination between superiors and inferiors. Item 10 has also a negligible correlation coefficient. Such items are usually not selected for the final inclusion in the test. Items 1, 2, 3, 6 and 9 have positive biserial *r* coefficient indicating that the items are discriminating. When all the item-total correlation coefficients have been obtained in the item analysis worksheet as shown in Table 4.1, they are ranked from highest to lowest. In general, the correlation coefficient of 0.20 or more is regarded as satisfactory and therefore, those items which have a validity index of 0.20 and above, are selected for final inclusion in the test. At this juncture, Nunnally (1970) has made a very useful suggestion regarding the selection of items. According to him, the selection of items should be directly dependent upon the level of reliability of the test desired. If the test constructor wants to have the reliability coefficient of the test as high as 0.80 and he has dichotomous items on hand, a sample of the first 30 items having the highest correlation (after they have been ranked from highest to lowest) may be selected. Subsequently the Kuder-Richardson (KR) formula 20 (this formula has been discussed in Chapter 5), on the basis of these 30 items, would be applied to examine whether or not the test has reached the desired reliability coefficient. If KR 20 yields a coefficient of 0.80, no other items are included in the test and the process of item analysis is finished. If KR 20 yields a coefficient below 0.80, a few items having the next highest correlations with total score should be added, and then KR 20 for the original 30 items plus the number of items added may be applied to compute the desired level of reliability.

When selecting bipolar items (which are common among the measures of personality, values and interests) on the basis of item-total correlations, sometimes interactive procedure becomes essential. Ordinarily, when at least 75% of the items yield positive item-total correlations, the test is said to be an ideal one. It is not uncommon that the remaining 25% items yield either zero or significant negative correlations. Interactive procedure is applied to those items yielding significant negative correlations so that they can now yield significant positive item-total correlations. In order to do this, scoring for all such items having negative correlations is reversed. For example, suppose a particular item of an extroversion test had been previously scored as +1 for "Yes" and had yielded negative correlation, now it would be scored as +1 for "No", and vice versa for items having the same negative correlations which previously had a score of –1 for "No". In this way, the scoring of all items yielding negative significant correlations with total score would be reversed. However, the scoring of items with positive correlation or nonsignificant negative item-total correlation would not be changed. Now, the entire test would again be administered to obtain a new set of total scores and for each item, new item-total correlations would be computed. It is expected this time that the number of positive item-total correlations will increase. If many more items having negative item-total correlations still remain, the procedure can be repeated.

A SIMPLIFIED ITEM ANALYSIS FORM

Item analysis may be done in still more simplified and convenient ways without using the above table. Suppose the investigator is constructing an intelligence test in which there are 50 items and each item has four options. For item analysis, the simplified procedure to follow would consist of the following steps.

1. Each item will be scored according to the scoring key which provides the correct answer and the correct answer will be valued as 1, 2 and so on, or by a weight decided by the investigator.

---

* For abac table prepared by Flanagan for computing the biserial *r* coefficients, students are advised to consult *Psychometric Methods* by J P Guilford, 1954, pp. 428–31. There, they can also find other abac tables for computation of point biserial coefficient, tetrachoric *r* and phi-coefficients.

2. The score given for each correct answer will be added and a total score for each examinee on that test will be determined.
3. The scored answer sheet will be arranged from high to low total scores.
4. Following the suggestion of Kelley (1939), the upper 27% and the lower 27% of examinees will be selected and the middle 46% will be left intact. The total *N* would be 370 so that the upper 27% will consist of 100 cases and the lower 27% will also consist of 100 cases.
5. The number of examinees in the upper 27% group who responded to each option (of Items No. 1 and 2, for example) is counted and entered in column 1 of Table 4.2. Similarly, in column 2 the number of examinees in the lower 27% group who responded to each option is entered.

Table 4.2 A simplified item analysis form

	Item	1 Upper 27% (*N* = 100)	2 Lower 27% (*N* = 100)	3 Difference	4 *D* or *V* (Discrimination Index)	5 Correctly responding from middle 46%	6 Total	7 *P* Difficulty Value
Item No. 1	A	0	0	60	+.60	110	210	.56 or .50
	B*	80	20					
	C	10	70					
	D	10	10					
Item No. 2	A	15	10	–20	–.20	130	230	.62
	B	20	20					
	C*	40	60					
	D	25	10					

* Correct alternative

6. The number of examinees in the lower group who selected the correct alternative is subtracted from the number of examinees selecting the correct alternative in the upper group, and is entered in column 3. In Table 4.2, the difference is 80 – 20 = 60. This is for item No. 1. The same value for item No. 2 is –20 where the number of examples selecting correct alternative in lower group exceeds the number of examinees selecting the correct alternative in upper group.
7. The difference found in column 3 (which is 60 in case of item No. 1) is divided by the number of examinees in upper 27% group (or lower 27% group) which is 100 here. Thus, $\frac{60}{100} = .60$ which is entered in column 4. The value of column 4 is called discrimination index, abbreviated as *D* or *V*. For item No. 2, the same value is – .20.
8. The number of examinees who responded correctly in the middle group (that is 46% which is 110 in case of item No. 1) is counted and entered in column 5. If the upper group and the lower group each contains half the group, there will, of course, be no middle group and column 5 will be omitted. The same value for item No. 2 is 130.

9. The number of examinees who responded correctly in the upper, lower and middle group is added. The sum of these three values represents the total number of examinees who answered the items correctly. The sum is 80 + 20 + 110 = 210 which is entered in column 6 for item No. 1. The same value for item No. 2 is 230.
10. The value entered in column 6 which is 210 in case of item No. 1, is divided by *N*, the total number of examinees taking the test. This is the proportion of the examinees in the total group who responded correctly (called difficulty value and often abbreviated as *P*). Thus, *P* here is $\frac{210}{370} = .56$ (by Equation 4.1) and if we apply Equation 4.2, *P* becomes $\frac{80+20}{100+100} = \frac{100}{200} = 0.50$. The same value for item No. 2 is .62.

Although the task of item analysis looks complex at first glance, it can be simplified and made convenient by following the above steps and by keeping records of the data wisely.

EFFECTIVENESS OF DISTRACTORS OR FOILS (DISTRACTOR ANALYSIS)

Item analysis, especially in multiple-choice items, aims at determining the effectiveness of the distractors. Indices of item difficulty and discriminative power simply indicate whether an item is doing its intended job and, not why or why not. If a particular item is not functioning well or is found to be defective, we have to examine its possible reasons and one obvious method is to examine the distractability of the incorrect options. This procedure is called *distractor analysis*. Usually, the number of examinees answering each distractor of an item in both the upper group and the lower group is counted. Ordinarily, any distractor to be called a good distractor must be answered by more examinees of the lower group. If this happens, a distractor is retained in the form it was presented in the item. But if a distractor is answered by more examinees of the higher group than the examinees of the lower group, the distractor is regarded to be of poor form, and the test constructor should rewrite or modify this distractor to make it appear more attractive to the examinees of the lower group. A test constructor may often find some nonfunctional distractors, which, by definition, are those that contribute nothing to the test. Such distractors are chosen by none of the examinees of either the upper group or lower group. Test constructors often *drop* such distractors from the item or carefully modify them to suit their purpose. There are also distractors which are equally attractive to both the upper and lower groups. Such distractors also need suitable modifications. Any unsuitable or poor distractor tends to lower the difficulty value of an item and, therefore, it must be carefully dealt with.

Analysis of distractors does not require application of any complex formula. Examination of the alternatives selected by the upper group and the lower group in the item analysis makes it obvious how the distractors worked. To be more sure and exact, we can compute the discriminating power or discriminating index of each distractor as it has been done in Table 4.2. On the basis of the discrimination indices of the distractors, that is A, C and D, it can be said that distractor C is working in the right direction because it attracted more examinees in the lower group. Distractors A and D are not working at all because A was chosen by none of the upper and the lower groups, and distractor D is chosen equally by both the groups. From these sets of information, we come to the conclusion that items and/or distractors need rewording. From this example, it is also obvious that the correct response should have positive discrimination index whereas distractors should have negative discrimination index.

If we don't have the extreme group method at hand, and working on the basis of the responses of all the examinees, we can compute the number of people expected to choose each of the distractors using the formula given on the next page.

$$\text{Number of persons expected to choose each distractor} = \frac{\text{Number of persons answering items incorrectly}}{\text{Number of distractors}} \quad (4.6)$$

Suppose, there are 200 examinees who answer to a particular four-option multiple-choice test. On item No. 22, for example, 110 responded to A, 30 to B, 45 to C and 15 to D (A here is the correct option). It means that 90 examinees answered the item incorrectly. We should, therefore, expect (by Equation 4.6) 30 examinees to choose each distractor. In fact, 30 examinees chose option B. More examinees chose C than expected, and fewer chose D. When the number of examinees choosing a distractor significantly exceeds the number expected, there have been reported two possibilities. First, it is just possible that the choice may reflect partial knowledge. A second possibility is that the item is poorly constructed. In a perfect test item, we might expect the responses of the examinees answering incorrectly to be divided equally among the number of distractors.

SPEED TESTS

Some standardized tests emphasize the speed of the response of the examinee. They provide a limited time within which an examinee is expected to reach the last item in the test. In other words, they place primary importance upon the speed with which an examinee can answer the item. Ideally, in a speed test all items should be of uniform degree of difficulty. Often item analysis of such speed tests provides a special problem to the test constructor. The index of difficulty and the index of discrimination cannot be determined in exactly the same way as is done in the case of a power test. As such, a separate discussion is needed.

Index of Difficulty

In a speed test many examinees never reach the items at the end of the test, not because of their intrinsic difficulty but because of their position in the test. Hence, the usual way to determine the index of difficulty by dividing the number of correct answers by the number of the examinees who take the test, would be a misleading index. Hence, in the speed test the index of difficulty of an item is determined by the number of correct answers to an item divided by the number of examinees who answer or reach that item (and not the total number of examinees). The formula for the index of difficulty becomes:

$$p = \frac{R}{N_r} \quad (4.7)$$

where p is the index of difficulty; R is the number of examinees who gave the correct answers; and N_r is the number of examinees who actually reached or answered (right or wrong) that item.

Suppose that a test consists of 50 items. Of these, the first 45 items were answered by all the examinees ($N = 100$) but the remaining five items were answered by only a few. Further, suppose that item 48 was answered by only 60 examinees out of which 20 answered it correctly. Likewise, item 50 was answered by only 20 examinees out of which only 5 examinees answered it correctly. Now, the index of difficulty for item 48 would be $= {}^{20}\!/_{60} = 0.33$ (and not $20/100 = 0.20$) and the index of difficulty of item 50 would be $5/20 = 0.25$ (and not $5/100 = 0.05$).

Like in a power test, the index of difficulty of an item is also corrected for chance success or guessing. The formula for the corrected index of difficulty of an item in a speed test is as under:

$$p_c = \frac{R - (W / K - 1)}{N - HR} \quad (4.8)$$

where p_c refers to the corrected proportion of the index of difficulty; R refers to the number of correct answers; W refers to the number of incorrect answers; K refers to the number of response options in the item; N refers to the number of total examinees, and HR refers to the number of examinees who could not reach the item within the time limit.

To illustrate, suppose an arithmetical test of 100 multiple-choice items (each having five options) was given to a group of 400 examinees who were required to answer them within a time limit of 30 minutes. Item 88 was correctly answered by 300 examinees, incorrectly answered by 60 examinees, and 40 examinees did not reach the item and hence, failed to attempt it within the time allowed. Now, its corrected index of difficulty would be equal to:

$$p_c = \frac{300 - [60/(5-1)]}{400 - 40} = \frac{300 - 15}{360} = \frac{285}{360} = 0.791 = 0.79.$$

The uncorrected index of difficulty for the same item by Equation 4.6 would be equal to $300/360 = 0.833$. Thus, guessing tends to inflate the difficulty value of the item in a way similar to that in a power test.

Index of Discrimination

The index of discrimination in a speed test cannot be determined through the correlation between the individual item score. Why? The answer is that in a speed test, the items are arranged in an arbitrary order and not from easier to difficult order as in most power tests. Besides, the time in the speed test does not allow all examinees to reach the last item in the test. Consequently, items placed in the beginning of the test will have a higher difficulty value whereas items placed towards the end of the test will have a lower difficulty value. Both these types of extreme items would not correlate well with other items of the test and hence, would also not correlate with the total score. Items placed in the middle of the test, on the other hand, will tend to correlate with other items and consequently, with the total scores. In a speed test there is no specification regarding the type of items to be placed at the beginning, middle or at the end of the test. Generally, items tend to appear in an order which is arbitrary as well as random. When the order is arbitrary, correlations of items with total scores also become arbitrary and there is no sense in choosing items on the basis of the item-total correlation.

In power test, the reliability of the test is dependent upon the number of items in the test. But this is untrue in case of speed tests where the reliability is dependent upon the reasonable time limit allowed for the completion of items. Therefore, in speed test items are selected according to the principles that will establish a reasonable time limit, which may yield a most reliable distribution of total scores. Nunnally (1970) has suggested a very useful method for establishing the time limit of a speed test. Suppose, the test constructor is preparing a test of numerical ability in which items tend to measure very elementary principles of addition and subtraction. Suppose the test constructor feels that five such items may be answered in one minute. Then, he may allot a total time of 20 minutes for 100 items. But this is simply a conjecture on the part of the test constructor. In order to establish the ideal time limit, he is required to do a little more. He may randomly select six groups of examinees having equal size, equal age, and equal intelligence. He may then give different time limits to each group for answering the same test. Thus, he may give 15, 22, 18, 25, 30 and 10 minutes to six groups respectively. The most ideal time limit would be one which will produce the most reliable distribution of scores. The reliability may be determined either by a suitable method of estimating the reliability coefficient or by the standard deviation of the total scores. It is convenient to get the estimate of reliability of such tests through the computation of standard deviation. The larger the standard deviation, the higher is the reliability of the test. Six time limits would yield six different standard deviations. The largest of these values of standard deviation will yield the highest reliable distribution of scores and consequently, that time limit will be the ideal time limit for the test. Suppose that the time limit of 25 minutes produced the largest standard deviation. Then, the ideal time limit of the sai numerical ability test would be 25 minutes and not 20 minutes as earlier decided by the test constructor.

Thus in a speed test, items are selected not on the basis of item-total correlation but on the basis of index of difficulty as well as upon ideal time limit experimentally determined for the test.

FACTORS INFLUENCING THE INDEX OF DIFFICULTY AND THE INDEX OF DISCRIMINATION

The difficulty of an item is influenced by several factors. The stem of the item may be complex and/or ambiguous and this may lower the index of difficulty of the item. Learning experience or previous experience of the examinee is also likely to affect the index of difficulty. If the examinee is not familiar with the type and content of items, he is likely to score poorly on the test. This will lower the index of difficulty. The nature of response alternatives also influences the index of difficulty. When the alternative responses are all homogeneous, the examinee is in a difficult position to choose the correct answer. This is true with multiple-choice items. When alternative responses are relatively less homogeneous, the examinee may easily hit the correct answer.

The index of discrimination is also influenced by several factors. Since the index of discrimination is intimately related to the index of difficulty, many factors which influence the index of difficulty also influence the index of discrimination. For example, previous experiences or learning experiences, extent of vagueness in an item and the homogeneity of alternative responses have a direct influence upon the index of discrimination by lowering it. Other factors influencing the index of discrimination are heterogeneity of the group of examinees, ability of the stem to frame a clear question, and the effectiveness of the distractors in appealing to those who do not know the correct answer.

PROBLEMS OF ITEM ANALYSIS

Despite the fact that test technicians have devised numerous methods of item analysis so that a test constructor may not face any difficulty or problem, some basic problems remain unsolved. These problems are enumerated below.

1. *Problem of spurious correlation in item-total correlation*: Whenever the total score is used as a criterion for validating the item, that is, whenever an individual item is correlated with the total score, the obtained correlation coefficients are spuriously high. The problem is more serious in the case of a homogeneous test where all items measure more or less the same function. This happens because in each total score, almost all the items make their contribution. In other words, each item is a part of the total score, which is used as the criterion in determining item-total correlation. Two general conditions may be specified for minimizing the extent of spurious correlation. The number of items should preferably be large in the test, and, as the items having 0.5 as index of difficulty have more of a spurious element than items of extreme indexes of difficulty, such extreme items should also be included in the test.

2. *Problem relating to dichotomous items*: Dichotomous items (or bipolar items) pose some intriguing problems in item analysis. Items with True-False, Yes-No, Agree-Disagree response options are common among nonability tests such as personality test, interest test, attitude scale, anxiety test, etc. Ability tests, however, do not emphasize the bipolar items and mostly concentrate upon the multiple-choice items. The problem in dichotomous items may be illustrated through an example. Suppose the test constructor has written 100 items for an anxiety test, each with Agree-Disagree response options. Fifty items are positive statements and 50 items are negative statements. The test constructor prepares the scoring key in a way that all agreements are scored as +1 and all disagreements are scored as zero. If such is the case, average correlation of each item with other items will be very close to zero and therefore, all item-total correlations will also be very close to zero. Moreover, all positive statements would correlate negatively with the negative statements. In a situation where the item-total correlation for all items approaches either near zero or becomes negative, items cannot be selected on the basis of the item-total correlation.

This problem of item analysis has one solution. The solution lies in the test constructor's ability to prepare the scoring key in such a way that it may yield most of the item-total correlations positive. In the above example instead of scoring all agreements as +1 and all disagreements as zero, a score of +1 may be given to all positive statements when agreed and to all negative statements when disagreed. Then, most of the item-total correlations will be positive, and items can be selected on the basis of the correlation of an item with the total score.

3. *Problem associated with control of unwanted factor*: Ideally, in homogeneous tests, all items should measure only one attribute and not any other attribute. Accordingly, items should correlate with each other and not with any other factor. But usually it happens that items also tend to correlate with a factor or factors, which are not wanted. For example, the items of a numerical aptitude test not only measure quantitative aptitude but also measure the factor of verbal comprehension because items also involve the comprehension of words and sentences used in items. In such a situation items, besides the factor of quantitative aptitude, would also tend to correlate with the unwanted factor of verbal comprehension. In an item-analysis situation, especially when items are to be selected on the basis of the item-total correlation, the effect of such uncontrolled and unwanted factors poses a problem—how to purify the items from such effects. Fortunately, there is a way out. The best procedure for the purpose of minimizing the effect is to start with four times the actual number of items to be retained in the final form. Thus, if a 40-item test is to be constructed, about 160 items are to be written. Each item then is correlated with the total score as well as with the unwanted factor. The purified items would be one which yield higher positive correlation with the total score and low correlation with unwanted factor. Items yielding low correlations with the total score and high correlation with unwanted factors are dropped. Subsequently, the Kuder-Richardson formula 20 may be applied to estimate the desired reliability of the test. If the reliability coefficient is higher than correlation of the total score with scores on the unwanted factor (which is computed separately), it is said that the effect of the unwanted factor in item selection on the basis of the item-total correlation has been minimized.

4. *Problem associated with guessing or chance success*: As discussed earlier, the problem o guessing or chance success is acute among the items with two-alternative options because he an examinee has 50% chances of answering correctly on the basis of a guess. In th multiple-choice items the problem of guessing is minimized, though not fully eliminated. Th effect of guessing is to elevate the score and thus, increase the index of difficulty a discriminative index. Test technicians, therefore, have argued for introducing penalty for ea guessing in form of correction. Generally, the penalty is introduced by subtracting the numbe wrong responses (attributable to blind guessing) from the number of correct responses. (formulas of correction have been discussed earlier). There are some test technicians who opp the idea of introducing correction for guessing on the ground that guessing does not necess reveal the lack of real knowledge being measured by the item. According to them, in many ca the formulas of correction for guessing do not make correction for guessing but for answe incorrect options due to misinformation or due to some other errors in construction of the itself. They further assert that if the test is sound, and gives sufficient time to every examin answer all items, and is appropriate for those who are taking it, no correction for guessing sh be introduced (because the chance for guessing is automatically minimized).

IMPORTANT INTERACTIONS AMONG ITEM CHARACTERISTICS

As we have seen, item analysis provides three types of information: (a) information about item difficulty (b) information about item discrimination (c) information about distractors. These three types of information are empirically related though conceptually distinct. It means that distractors affect difficulty value and difficulty value affects discriminating power. Likewise, distractors also affect discriminating power of an item. Here we shall see how these three characteristics interact with each other.

Distractors and difficulty: Any multiple-choice item could be very difficult or very easy depending upon its distractors. For example:

A. The first psychological laboratory was established by Wilhelm Wundt in the year:

(a) 1649

(b) 1779

(c) 1879

(d) 1959

B. The first psychological laboratory was established by Wilhelm Wundt in the year:

(a) 1865

(b) 1869

(c) 1889

(d) 1879

Even a person who knows little about the history of psychology would say that item B is difficult than item A because item B has plausible distractors whereas item A has ridiculous distractors. Obviously, the difficulty level of the item is being affected by the plausibility of the distractors. If the test-taker or the examinee knows nothing about the trait/domain being tested, any distractor might be equally plausible. But in reality, the test-taker knows the domain being tested and therefore, is not befooled by the ridiculous distractors. The reality is that the test-taker has generally an imperfect knowledge of the domain and therefore, may be easily befooled by the plausible distractors. Item difficulty can be easily changed by rewriting or modifying the same distractors.

Distractors and Discrimination: Distractors not only affect difficulty but also the discrimination value of an item. The presense of one or more implausible distractors tend to lower the difficulty of an item. Items which are too easy don't have any potential for making valid discrimination. More or less, the same is true for extremely difficult items. In view of this, item writers should pay special attention to writing good distractors because writing the stem part of the item or the correct answer is not very difficult. Thus distractors affect the level of discriminating power of an item via making the item either very difficult or very easy.

Difficulty and Discrimination: The level of difficulty value places a direct limit on the discriminating power of an item. If the item is such that every examinee chooses the correct answer ($p = 1.0$) or every examinee chooses an incorrect answer ($p = 0$), such an item response cannot discriminate between those who are superior with respect to the domain being measured and those who are inferior with respect to the domain being measured. When the difficulty value (or p value) is nearly 1.0 or 0, the ability of the item to discriminate between superior or inferior is negligible. Psychometricians have established that items with a p value near 0.50 such as 0.60, 0.70, 0.40 and 0.30 possess the maximum potential for making good discriminations. However, it should be kept in mind that a p value near 0.50 *does not guarantee* that an item will be a good discriminator. Extreme p value place a direct statistical limit on the discriminating power of an item. On the other hand when the p value is near 0.50, there are no statistical limits on the discrimating ability of an item. However, a poor item with a p value of 0.50 is still a poor item.

Thus, we find that all these three characteristics of items make sufficient interactions with each other.

THE ITEM CHARACTERISTIC CURVE (ICC) AND ITEM RESPONSE THEORY

The ICC summarizes much of the information conveyed by item analysis. It is a graphic representation of the probability of giving the correct answer to an item as a function of the level of the attribute assessed by the test. In fact ICC serves as a foundation of one of the most important theories, the item response theory. This theory was specifically developed for the purpose of understanding how individual differences in attributes of the examinees affect his behaviour when confronted with a specific item.

ICC is used to illustrate discrimination power and item difficulty. The steepness or slope of the ICC conveys information about the discriminating power of the item. When item-total correlation is positive, the slope of ICC is positive (curve A of Figure 4.1). When item-total correlation is near zero, the slope is near zero, that is, flat (curve B of Figure 4.1). When item-total correlation is negative, the slope of ICC is negative (curve C of Figure 4.1). Thus the relationship between the slope of ICC and item-total correlation gives an index of an item's discriminating power.

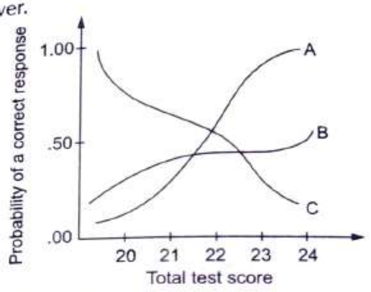

Fig. 4.1 Three ICCs showing discriminating power

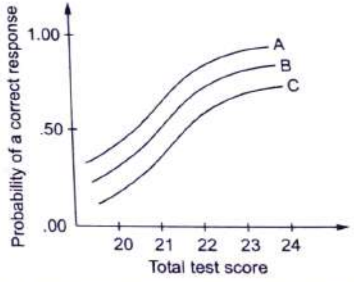

Fig. 4.2 Three ICCs showing variant difficulty values

The position of the ICC curve gives indication about the difficulty of each item. For the difficult items the ICC starts to rise on the right-hand side of the plot (high total test scores) and for the easier items, the ICC curve starts to rise on the left-hand side of the plot (low total test scores). Figure 4.2 shows ICC for three items which show similar discriminating power but which vary in difficulty. Item *C* is the most difficult and item *A* is the most easy one. Item *B* is of moderate difficulty.

Item response theory (IRT), also known as *latent trait theory* or *item characteristics curve theory*, is one of the modern approaches for explaining and analyzing the relationship between the characteristics of the examinee (ability) and responses to individual items. The theory states that the probability of a particular response to a test item is a joint function of one or more characteristics of the individual respondent and one or more characteristics of the test item. According to classical test theory, a score is derived from the sum of the individual's responses to different items which are sampled from a larger domain representing a trait or ability. IRT tends to consider the chances of getting particular items right or wrong. This new approach makes extensive use of item analysis (Cascio 1987; Steinberg & Thissen 1995). According to IRT each item on a test has an independent item characteristic curve that describes the probability of getting each particular item right or wrong, given a certain ability lavel of the examinee. In this new approach, with the help of computers, items can be sampled and the specific range of items where the examinee begins to have difficulty can be identified. In this way, the testers can make an ability judgement without subjecting the examinees to all the test items. Perhaps this is one of the most important advantages of IRT over the classical test score theory. That is the reason why IRT is considered as the most important development in psychological testing in the second half of the 20th century.

There are different approaches to the construction of tests using IRT. Most of them use two dimensions such as difficulty and discernability. Some other approaches add a third dimension for the probability that the examinee having a low level of ability will get a correct response. Whatever may be the approaches, they try to grade items in relation to the probability that those doing well or poorly on the test will have different levels of performances. Following Guion and Ironson (1983) item characteristic curves may be averaged to create a *test characteristic curve* which gives the proportion of responses expected to be correct for each level of ability.

In fact, IRT makes a set of assumptions about the mathematical relationship between a person's ability and the likelihood that the examinee will answer an item correctly. These assumptions form the basis of item characteristic curves regarding underlying relationships and the empirical outcomes of testing (Murphy & Davidshofer 1988).

One of the major advantages of IRT is that it provides measures that are generally *sample invariant*. In other words, it yields such measures for describing item characteristic curves that don't depend on the sample from which test data are drawn. Thus the theory produces measures of item difficulty, item discrimination, etc., that are invariant across different samples of people who take the test. This is not true for the classical standard item analysis statistic. For example, the same spelling test which is difficult for class IV students may be easier for class X students. Thus, the invariance of measures obtained using IRT is considered to be important since it allows characteristics of items to be analyzed without confounding them (as classical item analysis processes do) with the characteristics or abilities of the examinees.

The most attractive advantage of tests based upon IRT is that such tests can be easily adapted to computer administration. The computer can easily and rapidly identify the specific test items required to assess a particular ability level of the examinee. Thus, with this approach, the examinees don't have to suffer from embarrassment of attempting several items beyond their ability. Likewise, they are saved from wasting their time and efforts on items far below their abilities. In this approach, since every examinee gets a different set of items (which are apparently or likely to be matched with their abilities), the chance for cheating is greatly reduced. Weiss (1985) and Weiss and Yoes (1991) have readily demonstrated that computer-adaptive testing tends to increase efficiency by 50% or more by reducing the time required in responding to every item of the test. Additional advantages are that IRT can easily handle items that are written in different formats (Hayes 2000); it can identify those respondents who show unusual response pattern and it also reduces biases against people who are slow in completing the test items.

IRT has been applied to solve many such problems which are ordinarily difficult to be solved by traditional approaches. Three such examples are: (a) getting information about ability from distractors (b) tailored testing (c) item analysis for some specialized tests. These may be discussed as under.

(a) *Getting information about ability from distractors*: In classical item analysis format, an answer is either scored right or wrong. The right answer is given a score of 1 and the wrong answer is given a score of 0. It makes no difference which distractor (wrong answer) is chosen. Supporters of IRT believe that some distractors are, in fact, better answers than others because some of them reflect partial knowledge whereas others will be chosen only if the examinee has little knowledge. Thus on the basis of the responses given towards distractors, some inferences about the abilities of the examinee can be drawn. Thus IRT suggests that any response made towards a test item provides information about the person's ability or knowledge. Therefore, the theory puts emphasis upon the construction of a separate ICC for each possible response to a

multiple-choice item, that is to distractors as well as the correct response. Figure 4.3 presents a set of item characteristic curves for the possible responses to a single multiple-choice test item having three distractors *B*, *C* and *D* and one correct response, that is, *A*. Choice *A* shows positive discriminating power of the item.

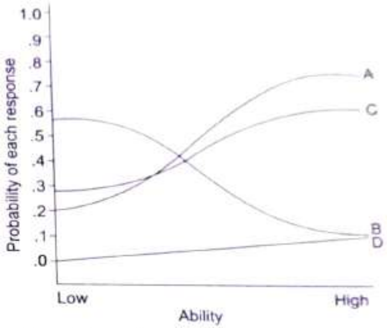

Fig. 4.3 Item characteristic curve for each response of a multiple-choice item

Choice C (one of the distractors) also shows positive discriminating power, choice B does not show any discriminating power and therefore, is unpopular and choice *D* shows negative discriminating power. Thus Figure 4.3 obviously suggests that although choice C possesses less discriminating power, it is nearly as good an answer as the choice C. Both *B* and *D* are unpopular choices. If a person has poor ability, it is very much likely that he will choose *B*. Thus rather than scoring the item simply either right or wrong (for all distractors), an examinee might be given nearly full credit for choosing *C* and might receive the lowest score for choosing B. In this way, a careful analysis of distractors might give some information about the ability of the examinees.

(b) *Tailored testing*: Another interesting application of item response theory is in computerized tailored testing. Today it is being realized that rather than constructing a long test that contains some items appropriate for the examinees having poor, moderate and higher ability levels, it would be better to construct tests that were tailored to the ability level of persons taking the test. Developments in microcomputer technology and in psychometric theory have led us to believe that for any individual, a test is most precise if the difficulty of items match with his ability level. Item response theory can be easily used to help tailor tests for the individual examinees. When an individual takes a test at a computer terminal, it is possible to estimate the examinee's ability at each step of testing and then to select the next test item which would match the person's estimated level of ability. Thus tailored testing provides a great deal of precision within a short test and IRT helps a lot towards this.

(c) *Item analysis for some specialized tests*: IRT is helpful in conducting item analysis of those tests where normal guidelines like item-total correlations should be positive, item difficulties clustered around .50, etc., fail to apply. Screening tests and criterion-keyed tests are the examples of such specialized tests where these guidelines don't apply. Screening tests are those that are used to make preliminary decisions or to make decisions as to whether or not the examinees exceed some minimum level of knowledge needed to qualify for the position. In criterion-keyed tests, there are no right or wrong answers and an examinee's test score is determined by the extent to which his responses become similar to the responses of some known group. The famous MMPI is an example of criterion-keyed test. Researches have shown that for analyzing these items of criterion-key tests as well as screening tests, IRT is of special use. With the help of IRT in such tests. Therefore, test users can easily determine the discrimination power and difficulty level of each item.

In this way, IRT has wider applications in the field of psychological testing. Despite the advantages of IRT, some technical problems have come up. For example, difficulties arise when the test measures multiple dimensions (Kaplan & Saccuzzo 2001).

☑ Review Questions

1. What is meant by item analysis? Discuss the methods of calculating difficulty value of items in a power test.
2. Discuss the major problems of item analysis and cite examples. Also, discuss the factors that influence item analysis.
3. What is meant by discrimination index? Discuss the major steps in calculating discrimination index and difficulty index of an ability test.
4. Distinguish between difficulty value and validity index of an item. Point out the importance of Kelley technique in item analysis.
5. Discuss the important interactions among various characteristics of an item.
6. What do you mean by Item Characteristics Curve? What is its significance in item analysis?
7. Examine the major implication of item response theory for test development.

◆

5

RELIABILITY

CHAPTER PREVIEW

- History and Theory of Reliability
- Meaning of Reliability
    - Logical Meaning or Technical Meaning of Reliability
- Methods (or Types) of Reliability
    - Test-Retest Reliability
    - Internal Consistency Reliability
    - Alternate-Forms Reliability
    - Scorer Reliability
- What is a Satisfactory Size for the Reliability Coefficient?
- Standard Error of Measurement
- Reliability of Speed Test
- Factors Influencing Reliability of Test Scores
    - Extrinsic Factors
    - Intrinsic Factors
- How to Improve Reliability of Test Scores?
- Estimation of True Scores
- Index of Reliability
- Reliability of Difference Score
- Reliability of Composite Score

HISTORY AND THEORY OF RELIABILITY

As compared to physical sciences, in psychology, many things make the task of measurement very difficult. One important reason for this is that psychologists are rarely interested in simple qualities like height, weight, width, etc. In fact, they are interested in measuring complex things/traits like intelligence, beauty, aggressiveness, etc., which can neither be seen nor touched. Besides this, no rigid yardsticks are available to assess such traits. In the absence of rigid yardsticks, the test developer must use "*rubber yardsticks*" which may sometimes be stretched to overestimate some measurements and shrunk to underestimate the same at other times (Nunnally & Bernstein 1994). Psychologists thus have to assess their measuring instruments to determine how much 'rubber' lies in them. This 'rubber' conceptualizes the 'errors' involved in the measurement. The theory of measurement error has, thus, been the major concern while psychologists attempt to work on reliability.

Psychology owes its debt to the British psychologist Charles Spearman who took pains to work towards the advanced development of reliability assessment. In 1733, De Moivre introduced the concept of sampling error (Stanley 1971) and in 1896 Karl Pearson had developed the popular product-moment method of correlation. Reliability theory tries to put these two basic concepts together in the context of measurement. Spearman's work provided the much-needed ntemporary reliability theory and his work was published in 1904 in an article entitled "The Proof and Measurement of Association between Two things". Spearman's contribution attracted the attention of the measurement pioneer E L Thorndike of the United States of America and in 1904, he published a book entitled *An Introduction to the Theory of Mental and Social Measurements* that contained a lot of such concepts which even by contemporary standards have been found to be very useful. Since 1904, many noted developments took place that further refined the process of the assessment of reliability. Among these, the most important was the article by Kuder and Richardson in 1937 in which several new reliability coefficients were introduced. Of these reliability coefficients, the coefficients obtained by Kuder-Richardson formula 20 and 21 became very popular later on. Cronbach and his colleagues made a major advancement (Cronbach 1989; Cronbach et al. 1972) by evolving various methods for identifying many sources of errors. More recently the item response theory (IRT) has been formulated and it has taken advantage of computer technology for advancing the basic concepts of psychological measurement (McDonald 1999; Michell 1999). However, the basis of IRT rests upon many of the ideas introduced by Spearman about 100 years ago.

The classical test theory assumes that any observed score (X) is equal to the true score (T) plus the error score (E). This classical theory assumes that each person has a true score that would be obtained if there are no errors in measurement. But in reality, the measuring instruments are not perfect. Such errors are also produced by the characteristics of the individuals or situations which can affect the test scores but which have nothing to do with the attribute being measured. The difference between the true score and the score obtained results due to their error of measurement, or error score. Or the same can be said in this way that the difference between the score the person obtains and the score he is really interested in, equals the error of measurement:

$$X - T = E$$

The classical theory of reliability further assumes that the errors of measurement are not systematic, but random. Using a rubber-yardstick terminology, it can be said that the classical test theory uses such rubber yardsticks in which the ruler stretches and contracts at random. The true score is obtained by finding the mean of observations from repeated applications (Kaplan & Saccuzzo 2001). When it is said that the measurement errors are essentially random, this does not mean that errors arise from random or mysterious processes, rather it means that the causes of errors in measurement are so varied and complex that these measurement errors act like random variables. If the errors have the characteristics of random variables then it is reasonable to assume that the errors are equally likely to be positive or negative, and that they are not correlated with true scores or with errors of other tests. In other words, it is assumed that:

Mean error of measurement = 0

True scores and errors are uncorrelated = 0

Errors on different measures are uncorrelated = 0

On the basis of these assumptions, a very extensive theory of test reliability has been developed (Gulliksen 1950; Lord & Novick 1968). Some important facts, too, have been derived from this theory that have obvious implications for measurement. One such fact is that this classical theory of reliability states that the variance of obtained scores is simply the sum of variance of true scores plus the variance of errors of measurement. In terms of equation,

$$O_x^2 = O_T^2 + O_E^2$$

where O^2 = variance or standard deviation squared. This formula, in effect, suggests that the test scores vary as a result of two factors: (a) variability in true scores, and (b) variability due to errors of measurement. If much of the variability observed in test scores can be accounted for in terms of error scores then the test scores will be inconsistent (poor degree of reliability). On the other hand, if the errors of measurement have little effect on test scores, the test scores will be consistent and

they will be reflecting those consistent aspects of performance which are labelled as true scores. The reliability coefficient provides an index of the relative impact of true and error scores on the obtained test scores.

The *domain sampling model* is another important concept in classical test theory. This model tries to deal with the problems created by using a limited number of items to represent a larger and complex ability/construct. Suppose a researcher wants to evaluate the spelling ability of a person. The ideal way would be to go systematically through a standard dictionary, enable the person to spell each word, and finally, determine the percentage of words correctly spelled. However, it is very unlikely that this would be done because no researcher would spend time for such an arduous task. Therefore, he would like to use a *sample* of words which will be used for evaluating the spelling ability. Had the researcher given all the words in the English language for evaluating the spelling ability, the correct percentage of words recalled would have been the 'true score'. In reliability analysis, the task of the researcher is to estimate how much errors are made in using the score from the shorter test as an estimate of the true ability. In fact, the domain sampling model conceptualizes reliability as the ratio of the variance of the observed score on a shorter test and the variance on the long-run true score. The measurement analyzed in this model is the error introduced by using a *sample of items* rather than the whole domain. As the researcher increases the number of samples, it represents the domain more accurately. As a result, the greater the number of items or samples, the higher the reliability. Reliability can be estimated from the correlation of the observed test score with the true score. Finding the true score is rarely possible. Therefore, it is assumed that when items are randomly drawn from a given domain, each test or group of items should provide an unbiased estimate of the true score.

The *generalizability theory* is another alternative approach presented by Cronbach et al. (1972) for measuring and studying consistency of test scores. The most serious weakness in the classical reliability theory is the concept of error measurement, which is assumed to be random in nature. Cronbach has pointed out that factors that determine the amount of measurement are different for internal consistency methods than for test-retest or alternate-forms methods. The reliability coefficient is taken as a ratio of the true score to true score plus the error score, but if the make-up of the true score and error score change when the estimation procedures change, something is definitely wrong.

The generalizability theory identifies both systematic and random sources of inconsistency that may contribute to the error score. Cronbach et al. (1972) have claimed that the classical reliability theory which breaks down the obtained test score into the true score and random error components, is simply a special case of the generalizability theory. The primary difference between these two theories is that the generalizability theory, unlike the classical theory, recognizes that error is not always random, rather there are some specific and systematic sources of inconsistency in measurement that must also be taken into consideration. In the generalizability theory, the focus is on our ability to generalize from one set of scores to a set of other plausible scores. The central question in the generalizability theory is "What are the conditions over which one can generalize?" or "Under what sorts of conditions would one expect either similar or different results than the ones obtained in one situation?" The generalizability theory, prompts us to ask questions like "In what situation is the test reliable?" rather than simply asking whether or not a test is reliable. According to this theory, a test can be reliable for some purposes but can be quite unreliable for other purposes. Thus, this theor basically focuses on what we wish to do with a test rather than on the test itself. Th generalizability theory claims that if the scores differ systematically according to when, where how the test was taken, these differences will affect the generalizability of the test scores as we as the meaning of the test reliabilit

Suppose we wish to assess an individual's need for sex using a projective test administered by an individual clinician. Since there is apprehension that the test results may be unreliable, we may have three different psychologists test the person and repeat the test three times with each psychologist. The resulting scores from such a procedure might look something like those depicted in Table 5.1. (Scores on this hypothetical test range from 1 to 100.)

Table 5.1 **Test scores for three different psychologists, each of whom tests the individual three times.**

Time of Testing	Psychologist A	B	C	Average
January	70	80	75	75
April	70	80	76	75.3
July	72	90	73	78.3
Average	70.6	83.3	74.6	

Now the question is: Is this test reliable? Strictly speaking, the answer depends upon how we define reliability. For example, both psychologists A and C have fairly consistent scores over time but B consistently differs from psychologists A and C. The three psychologists agreed best in January and April but agreed less in July.

Data presented in Table 5.1 can be explained by using the classical reliability theory, that is, we can treat inconsistencies in test scores as random errors. However, it will be more informative to examine the systematic differences between psychologists or between the various time periods and this information can be used in describing circumstances under which one can or cannot generalize test scores. ANOVA, a statistical technique, can be applied here for measuring the systematic effects of variables (psychologists, times, etc.) on the consistency of measurements. This statistic would provide estimates of the variability in test scores associated with systematic differences in scores assigned by psychologists and differences in scores obtained in January, April and July. Average test scores over the time periods of January, April and July were more or less similar. Therefore, variability due to different times of testing is negligible. On the other hand, there were consistent differences between psychologist B and psychologists A and C. As a result, a larger portion of variability in scores is associated with differences between psychologists. Lastly, a good deal of variability is not systematically associated with the psychologists or with the time of testing.

A simple look at Table 5.1 suggests that scores can be more authentically generalized over time (average scores at each time being more or less similar) than over psychologists (psychologists differ in terms of their average scores). Thus the question posed in the beginning, "Is this test reliable?", would be answered differently depending upon whether the stability of scores over time or the stability of scores obtained by different psychologists is wanted.

MEANING OF RELIABILITY

Reliability is one of the important characteristics of any test. In its simplest sense, reliability refers to the precision, or accuracy, of the measurement or score. A well-made scientific instrument should yield accurate results, both at present as well as over time. In other words, such an instrument should give consistent results. Reliability refers to this *consistency* of scores or measurement which is reflected in the *reproducibility* of the scores. When all other factors are held constant or somehow controlled, a reliable test is one that produces identical (or at least highly similar) results for an examinee from one occasion to the other. A test is said to be

consistent over a given period of time when all the examinees retain their same relative ranks of two separate testings with the same test. A test is also said to be consistent (when administered once) if the examinees who obtain high scores on one set of items also score high on an equivalent set of items, and those who obtain a low score on one set of items also score low on an equivalent set of items. The consistency of scores obtained upon testing and retesting is referred to as the 'temporal stability' of a test whereas consistency of scores obtained from two equivalent sets of items of a single test after a single administration is referred to as the 'internal consistency' of the test scores. Thus, reliability may be defined as the consistency of scores obtained from one set of measures to another. According to Anastasi & Urbina (1997, 85), reliability refers to "the consistency of scores obtained by the same individuals when re-examined with test on different occasions, or with different sets of equivalent items, or under other variable examining conditions." It is obvious from the above interpretation that reliability is never the property of the test itself, rather, it is the property of a test when it is administered to the examinees, that is, it is the property of the test scores. The correlation coefficient indicating temporal stability is known as the *coefficient of stability*, and the correlation coefficient indicating internal consistency is known as the *coefficient of internal consistency* or the *alpha coefficient*. Any statistical measure of reliability must indicate both the coefficient of stability as well as the alpha coefficient. For obtaining the coefficient of stability, the two sets of measurements or scores, found upon testing and retesting, are correlated with each other. Likewise, for obtaining the alpha coefficient, the two sets of measurements or scores by two equivalent sets of items of the same test after its single administration, are correlated with each other. This is why, statistically, reliability is also defined as the *self-correlation of the test*.

The meaning of reliability can be further clarified by noting the following important points:

1. Reliability is concerned with the test scores or results obtained with an assessment instrument and not with the test or instrument itself. In other words, reliability is the property of the test scores and not of the test itself. Any assessment instrument may have a number of different reliabilities, depending upon the situation in which it is used or on the group involved. Therefore, it is more appropriate to speak of the reliability of test scores or of the assessment results than of the test or assessment itself.
2. Reliability is a necessary but not sufficient condition for validity. A test that produces totally inconsistent results cannot provide valid information about the test scores. On the other hand, highly consistent test scores may be assessing or measuring the wrong thing or may be used in inappropriate ways. Thus, a low reliability is indicative of the fact that a low degree of validity is present but high reliability does not ensure a high degree of validity. In brief, reliability simply provides the consistency that makes validity possible.
3. Any estimate of reliability is always concerned with a particular type of consistency. It means that test scores (or assessment results) are not reliable *in general*. They are reliable over different periods of time, over different raters, over different samples of task, and so on. It is just possible for the test scores to be consistent in one of these respects but not in another.

Logical Meaning or Technical Meaning of Reliability

The reliability of a test is also defined from another angle. Whenever we measure something either in the social sciences or physical sciences, the measurement involves some kind of error. If the measurement is perfectly accurate, that is, free from all kinds of errors which are attributable either to the imperfection in the instrument or to the personality factors of the examinees being tested, the reliability will be perfect and the reliability coefficient will be +1.00. But this is an ideal goal which is rarely achieved in either the social sciences or in the physical sciences. Each measurement contains some error and therefore, reliability is never perfect. Thus, it can be said that each individual score or measurement is made up of two types of scores—the true score and the error score. In terms of an equation, it can be written as:

$$x_T = x_\infty + x_e \qquad (5.1)$$

where x_T = the actual obtained score; x_∞ = true score; and x_e = error score. (∞ is the sign of infinity which represents the true score.)

True score is the score which is free from errors occurring due to the chance factor as well as from other kinds of errors. It is indicated by the mean of a large number of scores made by the same person on the same test. Usually, a person's true score remains the same but his obtained score may vary from trial to trial because error score contributes to obtained score in each trial. Error score may be of two kinds: random (or chance) and systematic (or constant). *Random error score* sometimes inflates the score and sometimes depresses the score. Thus it works randomly in both the positive and negative directions. Chance errors, or random errors, are also known as the *errors of measurement*. In the long run, the positive and negative errors would tend to cancel each other and, therefore, the mean of all these errors of measurement would be zero. *Systematic errors* work constantly in one direction and, therefore, they would either tend to inflate or depress the score. The mean of such errors of measurement would not be zero. Let us take an example. Suppose we are weighing 10 girls who are each 16 years old. Now our weighing machine, due to some mechanical trouble, may sometimes give high readings or low readings of weight. If we take 100 readings (that is, each girl being weighed 10 times), chances are that the errors of high readings and low readings would cancel each other and the mean of these errors would be zero, thus leaving the true reading of weight alone. It may also happen that the weighing machine is giving high readings each time (or low readings each time), a case illustrating the systematic error. In this case, errors will go on accumulating on the positive side and the mean of such errors would not be zero. When such errors creep into the measurement, it becomes difficult to estimate the true score. When we administer a test, many factors are likely to contribute to the error score. Some of these factors are errors in scoring, errors in test administration, fluctuations in the examinee's motivational and emotional status, guessing, misunderstanding of instructions, misleading items, etc. The reliability is directly related to the size of the error score. The smaller the error score, the more reliable the test or the measuring instrument.

Since any obtained score is divided into the true score and the error score, the total variance of the test is also divided into components—the *true variance* and the *error variance*. Variance is defined as standard deviation squared or SD^2. In terms of an equation, the situation may be written as:

$$\sigma_T^2 = \sigma_\infty^2 + \sigma_e^2 \qquad (5.2)$$

where σ_T^2 = total variance of the test score; σ_∞^2 = variance of true score; and σ_e^2 = variance of error score.

Thus, the variance of the test score is equal to the variance of the true score plus the variance of the error score. Correlation between σ_∞^2 and σ_e^2 is usually zero because the correlation between x_∞ and x_e is assumed to be zero. In the modern test theory, the reliability of a set of test scores is logically defined as the *proportion of the true variance*. In other words, reliability is that part of the total variance which is the true variance. The proportions of the true variance and the error variance are found by dividing the true variance (σ_∞^2) and the error variance (σ_e^2) respectively by the total variance (σ_T^2). Thus the proportion of true variance = $\sigma_\infty^2 / \sigma_T^2$, and the proportion of the error variance = σ_e^2 / σ_T^2. Now, reliability coefficient (r_{tt}) by logical definition becomes:

$$r_{tt} = \frac{\sigma_\infty^2}{\sigma_T} \qquad (5.3)$$

According to Equation 5.3, reliability is the proportion of true variance. The reliability coefficient can also be computed in a substitute way where the coefficient becomes equal to one minus that part of the total variance which is the error variance. The substitute formula becomes:

$$r_{tt} = 1 - \sigma_e^2 / \sigma_T^2 \qquad (5.4)$$

Table 5.2 Hypothetical distributions of obtained, true and error score and their means, variances and standard deviations

Students	Obtained Score	=	True Score	+	Error Score
A	20		22		−2
B	14		12		+2
C	35		30		+5
D	25		28		−3
E	16		18		−2
Sums:	110		110		0
Means:	22		22		0
Variances:	58.369		43.191		15.178
SDs:	7.64		6.572		3.895

Table 5.2 shows the hypothetical relationship between *obtained, true* and *error* scores. For example, student A obtained a score of 20, had a true score of 22 and therefore, had −2 points of error. Likewise, student C had obtained a score of 35, a true score of 30 and therefore, had +5 points of error. In this table (Table 5.2), the true variance is 43.191 and the obtained variance is 58.369. The ratio 43.191/58.369 is 0.739 = 0.74, which is the reliability of these measurements. This means that 74% of obtained variance is attributable to true variance. If the ratio was 1.00, the obtained variance would be wholly attributable to the true variance. If the ratio were 0, there would be no relationship between the obtained scores and the corresponding true scores.

$$\text{By Equation 5.3: Reliability} = \frac{\text{True variance}}{\text{Obtained variance}} = \frac{43.191}{53369} = 0.739 = 0.74$$

$$\text{By Equation 5.4: Reliability} = 1 - \frac{\text{Error variance}}{\text{Obtained variance}} = 1 - \frac{15.178}{58369} = 1 - 0.260 = 0.74$$

Students should note that any test is neither perfectly reliable nor perfectly unreliable. Thus, reliability is not an absolute principle, rather, it is always a matter of degree.

METHODS (OR TYPES) OF RELIABILITY

There are four most common methods of estimating the reliability coefficient of test scores. These methods are: (i) Test-retest reliability, (ii) Internal consistency reliability, and (iii) Parallel-forms reliability, or Alternate-forms reliability, or Equivalent-forms reliability, or Comparable-forms reliability, and (iv) Scorer reliability. Test-retest reliability and parallel-forms reliability are the external consistency procedures because they compare findings from two independent processes of data collection with each other for verifying the reliability of the measure. A detailed discussion of each of these methods follows.

Test-Retest Reliability

In test-retest reliability, the single form of the test is administered twice on the same sample with a reasonable time gap. In this way, two administrations of the same test yield two independent sets of scores. The two sets, when correlated, give the value of the reliability coefficient. The reliability coefficient thus obtained is also known as the *temporal stability coefficient* and indicates to what extent the examinees retain their relative position as measured in terms of the test score over a given period of time. A high test-retest reliability coefficient indicates that the examinee who obtains a low score on the first administration tends to score low on the second administration, and its converse, when the examinee scores high on the first administration tends to score high on the second administration. In computing retest reliability, the investigator is often faced with the problem of determining a reasonable time gap between the two administrations of the test. There have been some disagreements among experts regarding the time lapse between the two administrations of the test. When the time is too short, it is likely to increase the reliability coefficient due to carryover and practice effect. If the time gap, on the other hand, is too long, it is likely to lower the reliability coefficient. The most appropriate and convenient time gap between the two administrations is a fortnight, which is considered neither too short nor too long. There are evidences to support that this time interval yields a comparatively higher reliability coefficient.

Test-retest reliability has its disadvantages. Factors relating to the disadvantages mainly contribute to the error variance of test scores. The test-retest method is a time-consuming method of estimating the reliability coefficient. This method assumes that the examinee's physical and psychological set-up remains unchanged in both the testing situations. But in reality, this is not so. In fact, the examinee's health, emotional condition, motivational condition and his mental set-up do not remain perfectly uniform. Not only this, the examiner's physical and mental make-up also changes. Besides, some uncontrolled environmental changes may take place during the administration of the test. All these factors are likely to make the total score of the examinee different from the first administration and thus, the examinee's relative position is likely to change, thereby lowering the reliability coefficient. Obviously, such factors contribute to the error variance and reduce the proportion of the true variance in total variance. In a nutshell, the source of error variance in the test-retest method is *time sampling*. Maturational effects also operate in contributing to the error variance. When the examinees are young children and the time interval between the two administrations is a comparatively long one, such effects are more obvious. Since maturational growth is not uniform for all young examinees, they are likely to produce a wider fluctuation in test score on the second administration, thus lowering the reliability coefficient of the test. Besides, when the examinee is once acquainted with types of items and their mode of answer, he is likely to develop a skill which may help him in the second administration. He is also likely to memorize many answers given in the first administration, especially if the second administration follows a week after the first one. All the acquired skill, knowledge and memory of the first answers are likely to help examinees in answering them in a more or less similar way the second time, thus helping them in retaining their same relative position. Obviously, these factors contribute to the true variance and are also likely to inflate the reliability coefficient of the test score. Apart from this, tests that measure constantly changing characteristics are not appropriate for test-retest evaluation. Despite all these limitations, the test-retest method is the most appropriate method of estimating reliability of both the speed test and the power test. For a heterogeneous test, too, the test-retest method is the most appropriate method of computing reliability.

Internal Consistency Reliability

Internal consistency reliability indicates the homogeneity of the test. If all the items of the test measure the same function or trait, the test is said to be a homogeneous one and its internal consistency reliability would be pretty high. The most common method of estimating internal consistency reliability is the *split-half method* in which the test is divided into two equal or nearly equal halves. The common way of splitting the test is the *odd-even method.* Almost any split can be accepted except the first half and the second half of the items. A division of this sort is not preferred because the nature of items in the two halves in a power test is different. Usually, the easier items are placed at the beginning or in the first half of the test and the comparatively difficult items are placed towards the end of the test or in the second half of the test. However, the odd-even method can be reasonably applied for the purpose of splitting. In this method, all odd-numbered items (like 1, 3, 5, 7, 9, etc.), constitute one part of the test and all even-numbered

items (like 2, 4, 6, 8, 10, 12, etc.) constitute another part of the test. Each examinee, thus, receives two scores: the number of correct answers on all odd-numbered items constitutes one score and the number of correct answers on all even-numbered items constitutes another score for the same examinee. In this way from single administration of the single form of the test, two sets of scores are obtained. Product moment (PM) correlation is computed to obtain the reliability of the half test. On the basis of this half-test reliability, the reliability for the whole test is estimated. The source of error variance in the split-half technique is *content sampling* or *item sampling* in which scores on the test tend to differ due to particular nature of items or due to differences in selection of items. Table 5.3 illustrates how a worksheet for odd-even reliability may be prepared.

Table 5.3 Worksheet for the odd-even reliability

Examinee	Number of correct score on odd-numbered items	Number of correct score on even-numbered items
A	16	15
B	14	13
C	18	12
D	12	9
E	19	18
F	8	7
.	—	—
.	—	—
Z	7	8

In Table 5.3, each examinee has two scores—one for the odd-numbered items and another for the even-numbered items. Subsequently, PM correlation is computed between two sets of scores. This amounts to reliability of the half test. When the reliability coefficient of the half test is known, the Spearman-Brown prophecy formula is used for estimating the reliability of the whole test. The formula is:

$$r_{tt} = \frac{2r_{{}^1\!/_2{}^1\!/_{11}}}{1+r_{{}^1\!/_2{}^1\!/_{11}}} \tag{5.5}$$

where r_{tt} is the reliability coefficient of the whole test; $r_{{}^1\!/_2{}^1\!/_{11}}$ is the reliability coefficient of the half test. Equation 5.5 may also be written as follows:

$$\text{Reliability of the whole test} = \frac{2 \times \text{reability of half test}}{1+\text{reliability of half test}}$$

To illustrate, suppose the reliability coefficient of half test $\left(r_{{}^1\!/_2{}^1\!/_{11}}\right)$ is 0.70. Then the reliability coefficient of the whole test (r_{tt}) by Equation 5.5 becomes:

$$r_{tt} = \frac{2 \times 0.70}{1+0.70} = \frac{1.40}{1.70} = 0.823 = 0.82$$

The advantage of the split-half method is that all data necessary for the computation of the reliability coefficient are obtained in a single administration of the test. Thus the variability produced by the difference in the two administrations of the same test is automatically eliminated. Therefore, a quick estimate of the reliability is made. That is why, Guilford and Fruchter (1973, 409) have described it as *on-the-spot reliability*. The disadvantage is that since both the sets of scores are obtained on one occasion, fluctuations due to changes in the temporary conditions within the examinee as well as due to the temporary changes in the external environment will operate in one direction, that is, either favourably or unfavourably, the obvious result of which would be either an enhancement or depression in the real coefficient of reliability. Another demerit of the split-half method is that it should not be used with a speed test.

If it is, the reliability coefficient is overestimated. A test can be divided into halves through different methods and, it has been found that each method yields a different coefficient of reliability. Undoubtedly, this is another weakness of the split-half technique of estimating the reliability coefficient of a test.

Rulon and Flanagan Formulas

An estimate of internal consistency is also made through the Rulon formula and the Flanagan formula. Both these formulas provide the reliability of the whole test (that is total-test score) and not the half test, and both formulas estimate the reliability coefficient on the basis of proportion of error variance in total variance of the test. The lesser the error variance, the higher the true variance and therefore, the higher will be the reliability.

The use of the Rulon formula requires that the test should be divided into two equal halves, either through the odd-even method or by any other method. Thus each examinee would have one subtotal score on odd-numbered items and another subtotal score on even-numbered items. A simple difference between the two subtest scores would indicate the error of measurement or chance error of each examinee and thus, would give an idea of the error variance. The Rulon formula is:

$$r_{tt} = 1 - \frac{\sigma_d^2}{\sigma_t^2} \qquad (5.6)$$

where r_{tt} = reliability coefficient; σ_d^2 = variance (that is, standard deviation squared) of the difference between two half scores for each examinee; and σ_t^2 = variance (or square of the standard deviation) of the total score. Total score for an examinee is the sum of his scores on the two halves of the test. Computation of reliability coefficient by the Rulon formula is illustrated in Table 5.4 in which an odd-even subscore of 10 examinees on a 12-item test has been given.

Table 5.4 Worksheet for the computation of reliability by the Rulon formula and the Flanagan formula

	1	2	3	4	5	6	7	8
No. of Examinees	Odd X_0	Even X_e	Difference $X_0 - X_e$ (d)	d^2	X_t	X_t^2	X_0^2	X_e^2
1	6	5	+1	1	11	121	36	25
2	6	5	+1	1	11	121	36	25
3	2	2	0	0	4	16	4	4
4	3	2	+1	1	5	25	9	4
5	4	3	+1	1	7	49	16	9
6	2	0	+2	4	2	4	4	0
7	4	4	0	0	8	64	16	16
8	6	6	0	0	12	144	36	36
9	5	6	−1	1	11	121	25	36
10	5	4	+1	1	9	81	25	16
	$\Sigma X_0 = 43$	$\Sigma X_e = 37$	$\Sigma d = 6$	$\Sigma^2 d = 10$	$\Sigma X_t = 80$	$\Sigma X_t^2 = 746$	$\Sigma X_0^2 = 207$	$\Sigma X_e^2 = 171$

$$\sigma_d = \frac{1}{N}\sqrt{N\Sigma d^2 - (\Sigma d)^2}$$

$$= \frac{1}{10}\sqrt{(10)(10) - (6)^2}$$

$$= \frac{1}{10}\sqrt{(100 - 36)} = 0.8$$

$$\sigma_d^2 = (0.8)^2 = 0.64$$

$$\sigma_t = \frac{1}{N}\sqrt{N\Sigma X_t^2 - (\Sigma X_t)^2}$$

$$= \frac{1}{10}\sqrt{(10)(746) - (80)^2}$$

$$= 3.25576 = 3.26$$

$$\therefore \quad \sigma_t^2 = (3.26)^2 = 10.627 = 10.63$$

Now, reliability coefficient by Equation 5.6 is

$$r_{tt} = 1 - \frac{0.64}{10.63} = 1 - 0.06 = 0.94$$

The Flanagan formula for estimating reliability is very similar to the Rulon formula. In the Flanagan formula, variance of the score of odd-numbered items and the score of the even-numbered items are calculated separately and then, an estimate of error variance is made. Thus, like the Rulon formula, it is not based upon the difference of the two half scores. The Flanagan formula is:

$$r_{tt} = 2\left(1 - \frac{\sigma_1^2 + \sigma_1^2}{\sigma_t^2}\right) \qquad (5.7)$$

where r_{tt} = reliability coefficient and σ_1^2 = variance of scores of the first half; σ_2^2 = variance of scores of the second half; and σ_t^2 = variance of the total score. The computation of the reliability coefficient by Equation 5.7 from the data in Table 5.3 may be illustrated as follows:

$$\sigma_1 \text{ of scores of the odd-numbered items} = \frac{1}{N}\sqrt{N\Sigma X_0^{\,2} - (\Sigma X_0)^2}$$

$$= \frac{1}{10}\sqrt{(10)(207) - (43)^2} = \frac{1}{10}\sqrt{2070 - 1849} = 1.49$$

$$\therefore \quad \sigma_1^2 = (1.49)^2 = 2.2201 = 2.22$$

$$\sigma_2 \text{ of the scores of even-numbered items} = \frac{1}{N}\sqrt{N\Sigma X_e^{\,2} - (\Sigma X_e)^2}$$

$$= \frac{1}{10}\sqrt{(10)(171) - (37)^2} = \frac{1}{10}\sqrt{1710 - 1369} = 1.85$$

$$\therefore \quad \sigma_2^2 = (1.85)^2 = 3.4225 = 3.42$$

Now, r_{tt} according to Equation 5.7 is given by

$$r_{tt} = 2\left(1 - \frac{2.22 + 3.42}{10.63}\right)$$

$$= 2\left(1 - \frac{5.64}{10.63}\right) = 0.939 = 0.94$$

Thus, the Rulon formula and the Flanagan formula have yielded the same coefficient of reliability from the data of Table 5.4, which automatically checks the accuracy of the computation. These two formulas are also applicable to the computation of reliability of the alternate forms of the test. The advantage of these two formulas over the Spearman-Brown formula is that one need not calculate the reliability coefficient of the half-test scores.

Kuder–Richardson Formulas and Coefficient alpha

Kuder and Richardson (1937) did a series of researches to remove some of the difficulties of the split-half method of estimating reliability. They were dissatisfied with split-half method of estimating reliability and therefore, they devised their own formulas for estimating the internal consistency of the test. Their formulas 20 and 21 have become very popular and well known. $K-R_{20}$ is the basic formula for computing the reliability coefficient and $K-R_{21}$ is the modified form of $K-R_{20}$. The main requirements for the use of the K–R formulas are:

(i) All items of the test should be homogeneous, that is, each item should measure the same factor or factors in the same proportion. In other words, the test should have inter-item consistency, which is indicated by high inter-item correlation. Thus, the test should be a unifactor one.

(ii) Items should be scored either as +1 or 0, that is, all correct answers should be scored as +1 and all incorrect answers should be scored as zero.

(iii) For $K-R_{20}$, items should not vary much in their indices of difficulty and for $K-R_{21}$, all items should be of the same difficulty value. If the indices of difficulty of items are not equal, the value of reliability yielded by $K-R_{21}$ would be substantially lowered from that computed from $K-R_{20}$. $K-R_{20}$ is:

$$KR_{20} = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_t^2 - \Sigma pq}{\sigma_t^2}\right) \tag{5.8}$$

where KR_{20} = reliability coefficient by K-R 20; n = number of items in the test; σ_t^2 = variance of scores on the test; p = proportion of correct answer to each item; q = proportion of incorrect answer to each item; hence it is equal to $1 - p$.

$K-R_{20}$ requires that the investigator must have the item analysis worksheet ready before him and only then can he compute the reliability coefficient. This is because the formula requires a knowledge of the difficulty value (proportion of correct answer) of each item. As such, the computation of the reliability coefficient through $K-R_{20}$ involves a considerable amount of work. Therefore, another formula has been suggested in which Σpq has been substituted by other terms. The use of $K-R_{21}$ does not demand the item analysis worksheet. The information needed for $K-R_{21}$ is the mean of total test score, *SD* of the total score and number of items in the test. The $K-R_{21}$ formula is, in fact, the simplification of $K-R_{20}$ arrived at by assuming that all items are of the same difficulty value. This is known as formula 21 which is expressed as:

$$KR_{21} = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_t^2 - n\bar{p}\bar{q}}{\sigma_t^2}\right) \tag{5.9}$$

It is obvious that Σpq of $K-R_{20}$ has been substituted for $n\bar{p}\,\bar{q}$ in $K-R_{21}$. $\bar{p}$ is equal to the average of correct proportion to each item and $\bar{q}$ is the average of the incorrect proportion to each item in the test. Other symbols are defined as usual. Statistically, $\bar{p} = \frac{M_t}{n}$ and $\bar{q} = \left(\frac{n-M_t}{n}\right)$ (or simply $1-\bar{p}$) where M_t is the mean of the total score and n is the number of items. K-R 21 may also be simplified as:

$$r_{tt} = \frac{n\sigma_t^2 - M_t(n-M_t)}{\sigma_t^2(n-1)} \tag{5.10}$$

where M_t is the mean of the total score on the test and other symbols are defined like those in Equation 5.8. A simple study of K–R 20 and 21 reveals that both these formulas are based upon item statistics like p, $\bar{p}$, q and $\bar{q}$. However, Equation 5.10 (a modified version of $K-R_{21}$) is based upon item total score statistics like mean of the total score and standard deviation of the total score. Equation 5.10 is very useful in estimating the reliability coefficient of the test in advance.

Table 5.5 illustrates the use of $K-R_{20}$ and $K-R_{21}$, where the item analysis worksheet showing the item score and total score of 10 examinees on a 15-item test has been given. (Correct response has been given a score of 1 and incorrect response a score of zero.)

$$M_t = \frac{\Sigma X}{n} = \frac{80}{10} = 8.0$$

$$SD = \frac{1}{N}\sqrt{N\Sigma X^2 - (\Sigma X)^2} = \frac{1}{10}\sqrt{(10)(796) - (80)^2} = 3.95$$

$$\therefore \quad \sigma_t^2 = (3.95)^2 = 15.6025 = 15.60$$

$$\bar{p} = \frac{M_t}{n} = \frac{8}{15} = 0.5333$$

$$\bar{q} = \frac{n - M_t}{n} = \frac{15-8}{15} = \frac{7}{15} = 0.4666$$

Now, the reliability coefficient by $K-R_{20}$ (that is, with Equation 5.8) will be:

$$r_{tt} = \left(\frac{15}{15-1}\right)\left(\frac{15.60 - 2.62}{15.60}\right) = \left(\frac{15}{14}\right)\left(\frac{12.98}{15.60}\right)$$

$$= (1.071)(0.832) = 0.891 = 0.89$$

The reliability coefficient from the same data by $K-R_{21}$ (that is, with Equation 5.9) will b

$$r_{tt} = \left(\frac{15}{15-1}\right)\left(\frac{15.60 - (15)(0.53)(0.46)}{15.60}\right) = \left(\frac{15}{14}\right)\left(\frac{15.60 - 3.657}{15.60}\right)$$

$$= \left(\frac{15}{14}\right)\left(\frac{11.943}{15.60}\right) = (1.071)(0.765) = 0.819 = 0.82$$

A simple approximation of $K-R_{21}$ is done through Equation 5.10, which estimates the reli coefficient as follows:

$$r_{tt} = \frac{15(15.60) - 8(15-8)}{15.60(15-1)} = \frac{234 - (8)(7)}{(15.60)(14)} = \frac{178}{218.4} = 0.815 = 0.82$$

Table 5.5 Worksheet for the computation of reliability by K–R formulas

Examinee	Item No.															X	X^2
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV		
1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	3	9
2	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	12	144
3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4
4	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	5	25
5	1	0	0	1	1	1	1	1	1	0	0	0	0	1	0	8	64
6	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	13	169
7	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	6	36
8	1	1	1	1	1	1	1	0	1	0	0	0	0	1	1	10	100
9	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	14	196
10	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	7	49
R	10	9	7	8	8	7	7	4	4	2	2	1	3	4	4		
W	0	1	3	2	2	3	3	6	6	8	8	9	7	6	6	ΣX	ΣX^2
p	1.0	0.9	0.7	0.8	0.8	0.7	0.7	0.4	0.4	0.2	0.2	0.1	0.3	0.4	0.4	$= 80$	$= 796$
q	0.0	0.1	0.3	0.2	0.2	0.3	0.3	0.6	0.6	0.8	0.8	0.9	0.7	0.6	0.6		
pq	0	0.09	0.21	0.16	0.16	0.21	0.21	0.24	0.24	0.16	0.16	0.09	0.21	0.24	0.24	$\Sigma pq = 262$	

It is obvious that if the reliability coefficient is computed from the same data by K–R$_{20}$ and K–R$_{21}$ or by its simple approximation formula, K–R$_{21}$ always underestimates the coefficient. Students should note carefully that data in Table 5.5 do not fulfil the exact requirements of the K–R formulas because items differ widely in difficulty value. If the split-half reliability were to be computed from this data it would, perhaps, be larger than the coefficient obtained by K–R$_{20}$. The underestimation of coefficient by K–R$_{20}$ (when compared to split-half reliability) would be due to the wider differences in difficulty values of all the 15 items.

The Kuder–Richardson formulas have such wide applicability in estimating the reliability coefficient of a test that many researches have been conducted to modify the formulas so that they may be used even in those situations where the requirement for the use of the formulas is least satisfied. Dressel (1940) has modified the K–R formula to apply to even those situations where items are differently scored. As we know, one of the requirements of the K–R formulas is that correct items should be given a score of +1 and incorrect items should be given a score of zero. Dressel's modified K–R formulas can be applied even to those situations where the correct answers are given different scores such as +2, +3, +1, etc., and wrong answers are given a score of zero. Dressel's formula is as follows:

$$r_{tt} = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_t^2 - \Sigma W^{2pq}}{\sigma_t^2}\right) \qquad (5.11)$$

where W = the weight or score assigned to each correct answer of the item; and other symbols are defined as those in Equation 5.8.

Likewise, Ferguson (1951) has further extended the K–R formulas to apply to situations where each item has more than two scores, e.g., +1, 0, –1, etc. Tucker (1949) has also provided a modification of K–R 20. The modified formula is very similar to K–R 21 with one additional term.

$$r_{tt} = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_t^2 - n\overline{pq} + n\sigma_p^2}{\sigma_t^2}\right) \qquad (5.12)$$

where σ_p^2 = variance of the proportion of correct answers to all items, which is statistically equal to:

$$\sigma_p^2 = \frac{\Sigma p^2}{n} - \bar{p}^2 \qquad (5.13)$$

All other symbols in Equations 5.12 and 5.13 are defined as those in Equation 5.9.

The K–R formulas have some limitations too. If the K–R formulas are used in a heterogeneous test where each item measures different functions or traits, they will underestimate the actual coefficient of reliability. Likewise, if they are used in a test where items differ widely in their indexes of difficulty, they would again yield a lower coefficient of reliability. Finally, with speed test or a test which is very close to a speed test, the K–R formulas should not be used.

Mathematically, it has been shown that K–R reliability coefficient is actually the mean of all split-half coefficients resulting from the different splittings of a test (Cronbach 1951). However, this is strictly true when the split-half coefficients are estimated by Rulon formula and not when they are found by correlation of halves and Spearman–Brown formula (Novick & Lewis 1967). The source of error variable in K–R coefficients is *content sampling*.

Coefficient alpha: The Kuder–Richardson formulas are applicable to the tests whose items are scored as 0 or 1 (right or wrong) according to some other all-or-none system. Some tests, however, may have multiple-scored items. We often see that on a personality inventory, testee or respondent receives a different numerical score on an item, depending upon whether he checks 'Sometimes', 'Usually' 'rarely' or 'never'. For calculating reliability of such tests, a generalized formula known as *Coefficient alpha or* α (also called Cronbach's alpha) has been formulated (Cronbach 1951; Kaiser & Michael 1975). It simply means that coefficient alpha estimates the internal consistency of the test in which items are not scored as 0 or 1 (right or wrong). The formula for coefficient alpha is

$$r_{tt} = \alpha = \left(\frac{n}{n-1}\right)\frac{\sigma_t^2 - \Sigma(\sigma_i^2)}{\sigma_t^2} \qquad (5.14)$$

Here meaning of the various terms of the formula are the same as those of Formula 5.8, except $\Sigma(\sigma_i^2)$. Formula 5.14 appears quite similar to K–R$_{20}$ formula. However, the difference is that Σpq has been replaced by $\Sigma(\sigma_i^2)$, the sum of variances of item scores. The procedure is to find out the variance of all individual scores on each item and then to add these variances across all items. Psychometicians today regard the coefficient alpha as the most general method of finding estimates of reliability through internal consistency. In fact, Cronbach's alpha tells us the overall consistency of the measure, that is, how much high responses go with highs and low responses with lows over all the items in the measure. In general, in psychological and educational research a good measure should have a Cronbach's alpha of at least .60 and preferbly closer to .90. Mathematically, Cronbach's alpha is the equivalent of average of all possible split-half reliability coefficients of the test. It ranges from zero (zero internal consistency) to 1 (perfect internal consistency) provided certain assumptions are met. A negative alpha coefficient indicates that the items of the test are negatively correlated and that an inappropriate reliability model has been used. The sources of error variance in coefficient alpha are *content sampling* and *conten heterogeneity*.

Alternate-Forms Reliability

Alternate-forms reliability is known by various names such as the *parallel-forms reliabili* *equivalent-forms reliability* and the *comparable-forms reliability*. Alternate-forms reliabili requires that the test be developed in two forms, which should be comparable or equivalent. T forms of the test are administered to the same sample, either immediately the same day or w the time interval of usually a fortnight. When the reliability is calculated on the basis of data collected immediately on the basis of two administrations of the test, it is called *alternate-form (immediate) reliability* and when the reliability is calculated on the basis of the data collected after a gap of a fortnight, it is called *alternate-form (delayed)* reliability. In the former, the source of error variance is *content sampling*, whereas in the latter, the sources of variance are *time sampling, content sampling* and *content heterogeneity*. Pearson *r* between two sets of scores obtained from two equivalent forms becomes the measure of reliability. Such a coefficient is known as the *coefficient of equivalence*. Alternate-forms reliability measures the consistency of the examinee's scores between two administrations of parallel forms of a single test. A very short time interval between the administrations of the two forms may help the examinees in maintaining the same position on the second form. This would apparently contribute to the true variance and would tend to raise the coefficient. Drastic changes in the content of the items of the two forms would contribute to the error variance and will tend to reduce the reliability coefficient of the test. A very long time interval is likely to produce demerits similar to test-retest reliability. The practice effect, memory effect including recall of items which occur in test-retest reliability due to identical contents of items in the two administrations are automatically controlled in parallel forms of the test because the second form is similar but not identical to the first form.

The biggest problem in parallel forms test is how to make both the forms equivalent in the true sense (because, if the test is not equivalent, the reliability coefficient may not be based upon the true variances). Gulliksen (1950) has defined parallel tests (or equivalent tests) as tests having equal means, equal variances and equal inter-item correlations. Freeman (1962, 72) has listed the following criteria for judging whether or not the two forms of the test are parallel.

1. The number of items in both the forms should be the same.
2. Items in both the forms should have uniformity regarding the content, the range of difficulty and the adequacy of sampling.
3. Distribution of the indexes of difficulty of items in both should be similar.
4. Items in both the forms should be of equal degree of homogeneity, which can be shown either by inter-item correlation or by correlating each item with subtest scores or with total test scores.
5. Means and standard deviations of both the forms should be equal or nearly so.
6. Mode of administration and scoring of both the forms should be uniform.

One can thus visualise the difficulty involved in making the two forms of a test parallel because it is very difficult, if not impossible, to meet all the above criteria. Moreover, such a test involves too much labour and time because all items in a changed language are written twice for the two separate forms. A test in order to be called a parallel test must meet the above criteria very closely, if not perfectly. Alternate-forms reliability is most appropriate when a speed test is being constructed. But this does not mean that it cannot be applied to a power test.

Scorer Reliability

There are tests such as tests of creativity and projective tests of personality which leave a lot to the judgement of the scorer. In fact, such tests are in as much need of scorer reliability as there is for the more usual reliability coefficients. Scorer reliability is the reliability which can be estimated by having a sample of test independently scored by two or more examiners. The two sets of scores obtained by each examiner are completed in the usual way and the resulting correlation coefficient is known as *scorer reliability*. This type of reliability is needed specially when subjectively scored tests are employed in research. The source of error variance in scorer reliability is interscorer differences.

Test-retest reliability, internal consistency reliability and parallel-forms reliability express reliability in terms of the correlation coefficient. As we know, the correlation coefficient is always relative; the reliability obtained by the said methods is, therefore, also known as the *relative reliability*. Sometimes analysis of variance (ANOVA) is also applied as a measure of relative reliability. Analysis of variance technique has been used by Hoyt (1941), Jackson (1939) and Alexander (1947). In applying the analysis of variance technique to individual items and determining the error variance thereupon, Hoyt has made four assumptions. The first assumption is that the total score of an examinee on a test can be divided into four independent components: (a) a component which is common to all examinees and to all items of the test; (b) a component associated with items only; (c) a component associated with the examinees only; and (d) the error component independent of the first three factors. The second assumption is that the variance of the error component of each item is equal. The third assumption is that the error component for each item is symmetrical and usually distributed; and lastly, the error component of any two distinct items is independent. Hoyt's formula for the reliability coefficient is:

$$r_{tt} = 1 - \frac{V_r}{V_e} \tag{5.15}$$

where r_{tt} = reliability coefficient; V_r = error variance; and V_e = variance among examinees.

It has been demonstrated by many research workers that Equation 5.15 yields a reliability coefficient which becomes identical with that obtained from K–R$_{20}$. And, therefore, some of the limitations of K–R$_{20}$ automatically apply to Hoyt's formula. For example, Hoyt's formula should not be used in a test where speed is an important factor. The analysis of variance approach for estimating the reliability coefficient can also be applied to data obtained from alternate forms of the test and retest administrations of the single form of a test.

WHAT IS A SATISFACTORY SIZE FOR THE RELIABILITY COEFFICIENT?

There is no hard and fast rule to determine the satisfactory level of the reliability coefficient of a test. In fact, whether the obtained coefficient of correlation is satisfactory or not, is judged by the purpose for which the test is given to the examinees. If the purpose is to segregate superiors and inferiors (that is, to make individual diagnosis), as is true with intelligence tests, aptitude tests and achievement tests, the reliability coefficient of 0.90 or higher is regarded as the best coefficient. Likewise, where the purpose of the test is to compare the means of the two groups of narrow range, a reliability coefficient of 0.50 or 0.60 would suffice.

High levels of reliability, in general, are most necessary when (a) tests are administered for taking a final decision about people (b) persons are categorized upon relatively small individual differences. For example, tests are often used for placement in the armed forces. If the tests used in the placements are unreliable, the decisions made regarding thousands of persons will also be unreliable.

Lower levels of reliability are acceptable when

(a) tests are used for preliminary rather than final decisions, and

(b) tests are used for categorizing people depending upon gross individual differences.

There are sometimes practical and theoretical obstacles in maximizing the reliability of tests. The practical obstacles may be exemplified with a fact that a reliability of 0.90 might be twice expensive to develop in comparison to a test with a reliability of 0.85. Thus the gain in reliability might not be worth the cost incurred. There are also some theoretical reasons why it is impossible to achieve high levels of reliability in measuring a particular trait/ability. In fact, there is an inevitable conflict between the goal of attaining precision (reliability) in measurement and the goal of attaining the breadth. This is known as *Bandwidth-Fidelity dilemma* (Cronbach & Gleser

1965; Shannon & Weaver 1949). These two terms, that is, Bandwidth and Fidelity have been taken from communication theory. Bandwidth refers to the amount of information contained in a message, while fidelity refers to the accuracy with which the information is passed. In general, the greater the amount of information to be conveyed (bandwidth) the lesser the level of accuracy desired, and less information can be conveyed. In a similar vein, bandwidth and fidelity also apply to the field of testing. For any test, the researcher has to decide between measuring a very specific attribute with high degree of accuracy or measuring broader attribute with lesser degree of accuracy. For example, suppose the researcher constructs a 50-item test of history for assessing the knowledge of history in the 15th century. On the other hand, he constructs the same 50-item test for assessing knowledge of history from ancient to modern period. In the former case there would be a highly reliable measure because bandwidth is limited and fidelity will be high whereas in the latter case, the test would be able to measure each student's general knowledge of history but would not be able to obtain a very accurate or reliable measure (poor fidelity) of such a broadly defined area (greater bandwidth).

STANDARD ERROR OF MEASUREMENT

So far, we have discussed only relative reliability which is expressed in terms of the coefficient of correlation. Many researchers prefer to express the reliability of test scores in terms of a new statistic called *standard error of measurement* rather than in terms of the reliability coefficient. As we know, the reliability coefficient is affected by the variability in the range of scores. The greater the variability in the total scores, the higher the standard deviation and therefore, the higher the reliability coefficient. A homogeneous set of test scores tends to produce lower reliability. One of the features of standard error of measurement is that it is not influenced by the variability in the range of scores. Standard error of measurement is almost the same even when the total scores of a sample vary widely. Statistically, standard error of measurement is defined as the standard deviation of the error component (or score) in the obtained test scores. Standard error of measurement may also be defined from the angle of true score and then, standard error of measurement becomes standard deviation of a sample of scores of an examinee about his true score. For calculating standard error of measurement, one must know either the error score or the true score of each examinee. If one can find out the error score (or error of measurement) of each examinee's test score, it is possible to calculate the standard deviation of the error score and the resulting statistic would be known as the standard error of measurement. But there is no way of calculating the error score directly and as such the standard error of measurement cannot be calculated directly. If the same test is repeated to the same group of examinees on unlimited number of times, the means of all such obtained scores for each examinee will be their true scores. Subsequently, the standard deviation of these scores (on that test) could be calculated and the resulting statistic would be known as the standard error of measurement. But again the problem is that it is not possible for anyone to repeat the same test on the same group of examinees on unlimited number of times, and hence the true score cannot be determined and therefore, the standard error of measurement cannot be statistically determined.

Standard error of measurement is calculated indirectly on the basis of the standard deviation of the test scores and the reliability of the test. The formula for standard error of measurement is:

$$SE_{meas} = \sigma\sqrt{1-r_{tt}^2} \tag{5.16}$$

where SE_{meas} = standard error of measurement; σ = standard deviation of the test scores; and r_{tt} = the reliability coefficient of the test calculated by any method of estimating the reliability.

To illustrate, suppose, the reliability coefficient of a test is 0.90 and the standard deviation of the test is 10, then substituting them in Equation 5.16, the $SE_{meas} = 10\sqrt{1-(0.90)^2}$ $= 10\sqrt{1-0.81} = 10\sqrt{0.19} = (10)(0.435) = 4.35$. Now, suppose that an examinee's obtained score on a test is 60. Interpreting this score, it can be said that chances are 95 in 100 (5% level of significance) that this obtained score is not more (or less) than 8.53 units (4.35×1.96) from his true score, that is, only 5 scores out of 100 scores (if the same examinee is given the same test 100 times) would be either less than 51.47 or greater than 68.53. Similarly, the chances are 99 in 100 (1% level of significance) that the obtained score will not be more (or less) than 1122 (4.35×2.58) units from his true score. Thus standard error of measurement provides a direct indication of the absolute accuracy of test scores. In all these cases, the true score is regarded as the mean of a repeated number of scores of the same individual on the same test.

If the reliability coefficient of a test is 1.00 (which is rare), the standard error of measurement will be zero because the right side of the equation is reduced to zero (cf. Equation 5.16). It should be noted that the smaller the standard error of measurement, the more reliable or consistent are the test scores.

RELIABILITY OF SPEED TEST

The distinction between speed and power test is difficult to be drawn in actual practice. In fact, this distinction is one of degree and most tests depend upon both power and speed in varying proportion.

Single trial reliability coefficients like those found by Odd-even or Kuder–Richardson techniques are inapplicable to speed tests. In speed test the individual differences in test scores are dependent upon the speed of performances and as such, the reliability coefficients found by these methods will be spuriously inflated. This can well be illustrated through an example: suppose a 100-item test is completely based on speed so that individual differences in scores are based wholly on the number of items attempted, rather than upon any other kinds of errors. In such a situation if A obtains a score of 80, he obviously will have 40 correct odd items and 40 correct even items. Likewise, individual B with a score of 50, will have odd and even scores of 25 and 25 respectively. Consequently, the correlation between odd and even scores would be perfect or + 1.00 which will be entirely spurious and provides no correct information about reliability.

Now the question arises: What alternatives are available for estimating the reliability of a speed test? As an answer, some suggestions can be made like this: Test–retest method, if applicable, can be applied. Equivalent-form reliability can also be properly employed. Split-half techniques can also be applied, provided the split is made in terms of time rather than in terms of items (Anastasi & Urbina 2002). In other words, such a split must be based upon separately timed parts of the test. The obvious way of splitting such test is to administer two equivalent halves of the test with separate time limits. For example, all odd items and all even items may be printed on separate pages and each set of items be given one-half the time limit of the entire test. Subsequently, either the Spearman–Brown prophecy formula or some other appropriate formula can be used to find out the reliability of the whole test.

If due to any reason it is not feasible to administer two half-tests separately, an alternative technique is to divide the total time into quarters and to find out a score for each of the four quarters. Subsequently, the number of items correctly completed within the first and fourth quarters can then be combined to represent one half-score while those in the second and third quarters may be combined to yield the other half-score. Such a combination of quarters also tends to balance the cumulative effects of practice, fatigue and other factors. This method is especially satisfactory when items don't differ sharply in difficulty level.

FACTORS INFLUENCING RELIABILITY OF TEST SCORES

The reliability of test scores is influenced by a large number of factors and all these factors can be categorized under two heads: extrinsic and intrinsic. *Extrinsic factors* are those factors which lie outside the test itself and tend to make the test reliable or unreliable. For example, variability in the range of ability of a group, environmental conditions, guessing by the examinee, momentary fluctuations in the individual, etc., constitute examples of extrinsic factors. *Intrinsic factors*, on the other hand, refer to those factors which lie within the test itself and influence the reliability of the test. Characteristics of items, total score, length of the test are some of the factors which constitute examples of intrinsic factors. A detailed discussion of both these factors follows.

Extrinsic Factors

Important extrinsic factors affecting the reliability of a test may be enumerated as follows:

1. *Group variability*: When the group of examinees being tested is homogeneous in ability, the reliability of the test scores is likely to be lowered. But when the examinees vary widely in their range of ability, that is, the group of examinees is a heterogeneous one, the reliability of the test scores is likely to be high. The effect of variability on reliability can be examined by seeing what happens when variability is zero. In such an extreme case all individuals receive the same score and therefore, standard deviation is zero. Each person is then at the mean of his group and has z score of zero. Since correlations (and the reliability coefficients) are defined as the average product of z scores, the product of z scores and the reliability coefficient in this situation are both zero. Only when there is some variability in the group, correlation and reliability are possible.

2. *Guessing by the examinees*: Guessing in a test is an important source of unreliability. In two-alternative response options, there is a 50% chance of answering the items correctly on the basis of the guess. In multiple-choice items, the chances of getting the answer correct purely by guessing are reduced. Guessing has two important effects upon the total test scores. First, it tends to raise the total score and thereby makes the reliability coefficient spuriously high. Second, guessing contributes to the measurement error since the examinees differ in exercising their luck over guessing the correct answer. Two examinees guessing on 60 `Yes-No' items out of 100, may differ—one examinee may get 80% of the 60 items correct purely by guessing and another examinee may get only 30% of 60 items correct purely by guessing. Such differences would make a difference in the obtained score by contributing more to the error score (or error of measurement) than to the true score. The higher the contribution to the error score, the lower is the reliability of the test.

3. *Environmental conditions*: As far as possible, the testing environment should be uniform. Arrangement should be such that light, sound, and other comforts are equal and uniform to all the examinees, otherwise it will tend to lower the reliability of the test scores.

4. *Momentary fluctuations in the examinee*: Momentary fluctuations influence the test score sometimes by raising the score and sometimes by lowering it. Accordingly, they tend to affect the reliability. A broken pencil, momentary distraction by the sudden sound of an aeroplane flying above, anxiety regarding noncompletion of homework, mistake in giving the answer and knowing no way to change it, are some of the factors which explain momentary fluctuations in the examinee.

Intrinsic Factors

The main intrinsic factors affecting the reliability of a test are as follows:

1. *Length of the test*: A longer test tends to yield a higher reliability coefficient than a shorter test. Lengthening the test or averaging total test scores obtained from several repetitions of the same test tends to increase the reliability. It has been demonstrated that averaging the test scores of several applications essentially gives the same result as increasing the length of the test. For example, suppose the test scores are being averaged after its three repeated applications and the test is lengthened three times the present length and administered once, then statistically the result of both averaging and lengthening will be the same. Care has to be taken to see that added

items should have the same variance and the same inter-item correlations as items of the original test. When the test has been increased or deceased by any multiple (that is 3, 4, or 5 times the original length), the Spearman–Brown formula may be used to estimate the reliability of the test. The formula is:

$$r_{nn} = \frac{(n)(r_{tt})}{1+(n-1)r_{tt}} \tag{5.17}$$

where r_{nn} = reliability coefficient of lengthened test; n = number of times the test has been increased; r_{tt} = reliability coefficient of the original test.

To illustrate, suppose an intelligence test of 100 items has a reliability coefficient of 0.80. If the test is increased four times its present length, that is, 300 more items are added so that now the test becomes 400 items, its reliability by Equation 5.17 becomes:

$$r_{nn} = \frac{(4)(0.80)}{1+(4-1)(0.80)} = \frac{3.2}{1+3\times 0.80} = \frac{3.2}{1+2.4} = \frac{3.2}{3.4} = 0.94$$

Thus, increasing the length four times the original or repeating the same test on the same examinees four times increases the reliability coefficient from 0.80 to 0.94. Thus the reliability of the test increases with an increase in length, though not in the same proportion. In case the total number of items is decreased from 100 to 50, *n* will be 1/2.

There is another Spearman–Brown formula through which it is possible to estimate the length of the test required to achieve a given level of the reliability coefficient. Again, the length of the test can be estimated in any multiple (that is, 2, 3, 4, or 5 times). The equation is:

$$n = \frac{r_{nn}(1-r_{tt})}{r_{tt}(1-r_{nn})} \tag{5.18}$$

where n = number of any multiple by which the test is to be lengthened; r_{nn} = level of reliability coefficient required; r_{tt} = reliability of the existing test.

To illustrate, suppose the reliability of an intelligence test is 0.60. For how much time should the test be lengthened in order to reach a reliability coefficient of 0.90? The answer can be given by solving the equation given in Equation 5.18 as mentioned below:

$$n = \frac{0.90(1-0.60)}{0.60(1-0.90)} = \frac{(0.90)(0.40)}{(0.60)(0.10)} = \frac{0.36}{0.06} = 6$$

(When *n* is in fraction, it should be rounded to the next whole number such as 2.3 = 3, and so on.)

Thus, in order to achieve a reliability coefficient of 0.90, the present test should be lengthened six times its present length. The use of both the above Spearman–Brown formulas makes two assumptions. The first assumption is that the new items added to the original test must have the same statistical properties as the items of the original test. In other words, the added item must have the same average difficulty value and it must have the same inter-item correlation among the items. The second assumption is that the added items must not influence the examinees' response.

According to Spearman–Brown formula, if a test initially has a reliability of 0.50, doubling the number of items increases the coefficient of reliability to 0.67. If the length of the test is again doubled, the estimated reliability increases to 0.80. Ebel (1979) has shown that doubling the length of a test quadruples true variance while it only doubles the error variance. Suppose that a test of thirty items has an obtained variance of 40 and true variance of 20, and thus reliability by Equation 5.3 will be $^{20}/_{40} = 0.50$. Now, further suppose that the number of items is doubled. According to Spearman–Brown formula (Equation 5.16), the reliability for a sixty-item test would be 0.67 if the thirty-item test had a reliability of 0.50. According to Ebel, because the length of the test was doubled, the true variance increases fourfold. Since the true variance was 20, it now becomes 20 × 4 = 80 whereas error variance only doubles. The error variance was 40 − 20 = 20. This error variance only doubles, that is, 20 becomes 40. Thus obtained variance becomes 80 + 40 = 120. Then reliability by Equation 5.3 becomes $^{80}/_{120} = 0.67$, the same value that would have been computed from Spearman–Brown formula.

2. *Range of the total scores*: If the obtained total scores on the test are very close to each other, that is, if there is lesser variability among them, the reliability of the test is lowered. On the other hand, if the total scores on the test vary widely, the reliability of the test is increased. Putting this in statistical terms, it can be said that when the standard deviation of the total scores is high, the reliability is also high and when the standard deviation of the total scores is low, the reliability is also low.

3. *Homogeneity of items*: Homogeneity of items is an important factor in reliability. The concept of homogeneity of items includes two things—item reliability (or inter-item correlation) and the homogeneity of function or trait measured from one item to another. When the items measure different functions and the intercorrelations of items are zero or near it (that is, when the test is heterogeneous one), the reliability is zero or very low. When all items measure the same function or trait and when the inter-item correlation is high, the reliability of the test is also high.

4. *Difficulty value of items*: In general, items having indexes of difficulty at 0.5 or close to it, yield higher reliability than items of extreme indexes of difficulty. In other words, when items are too easy or too difficult, the test yields very poor reliability (because such items do not contribute to the reliability)—poorer than when items are of moderate difficulty values. But sometimes such items could also be wholly indiscriminatory and then they would contribute nothing to reliability (cf. Table 4.1).

5. *Discrimination value*: When the test is composed of discriminating items, the item-total test correlation is likely to be high and then, the reliability is also likely to be high. But when items do not discriminate well between superior and inferior, that is, when items have poor discrimination values, the item-total correlation is affected, which ultimately attenuates the reliability of the test.

6. *Scorer reliability*: Scorer reliability (also known as reader reliability) is also an important factor which affects the reliability of the test. By scorer reliability is meant how closely two or more scorers agree in scoring or rating the same set of responses. If they do not agree, the reliability is likely to be lowered.

HOW TO IMPROVE RELIABILITY OF TEST SCORES?

Reliability of test scores can be improved by controlling those factors which adversely affect the reliability of the test. The following suggestions are useful for improving the reliability:

1. The group of examinees should be heterogeneous, that is, the examinees should vary widely in their ability or trait being measured.
2. Items should be homogeneous.
3. Test should preferably be a longer one.
4. As far as possible, items should be of moderate difficulty values; in other words, the indexes of item difficulty should have the range of 0.40–0.50–0.60.
5. Items should be discriminatory ones.

Apart from these general suggestions, there are three common approaches for improving the reliability of the test. One approach emphasizes upon the length of the test; the second approach emphasizes upon throwing out items that pulls down the reliability; and the third approach emphasizes upon correction for attenuation.

The approach emphasizing upon increasing the length of the test assumes that if new items similar to the original set of items are added, the reliability of the test would tend to increase. Following domain-sampling model, each item in the test is an independent sample of the trait or ability being measured. The larger the sample, the more likely the fact that the test will represent the true characteristic. According to this model, the reliability of a test increases as the number of items increases. Formula 5.18 is used to estimate the *n* times the test should be lengthened in order to reach the projected level of reliability.

Another approach to improve reliability is to discard the items that run down the reliability. Under this approach, two techniques are commonly applied: *factor analysis* and *item analysis*.

For increasing reliability, it must be ensured that all items measure the same thing. Factor analysis ensures that tests are most reliable if they are unidimensional (Loehlin 1998; Tabachnick & Fidell 1996). This reflects that one factor should account for considerably more of the variance than any other factor. Items that fail to load on this factor should be omitted or discarded. In item analysis, generally the correlation between each item and the total score for the test is examined. This type of item analysis is called *discriminability analysis*. When the correlation between a single item and total test score is low, the item is probably measuring something different from other items on the test. On the other hand, it may also mean that the item is so easy or so hard that the sample does not differ in its response to it. In either case, the low correlation indicates that such items are pulling down the estimate of reliability and therefore, they should be excluded.

The third approach to enhance reliability is to go for correction for attenuation for which reader is referred to consult chapter 6.

ESTIMATION OF TRUE SCORES

True score is statistically defined as the mean of an unlimited number of measurement or scores obtained by the same examinee on the same test. It is practically impossible to take an unlimited number of measurements or scores. As such, experts have devised an indirect method of estimating the true score of an examinee. When the reliability of the test is known, an examinee's true score can be estimated by a regression equation. The equation is as follows:

$$\overline{X}_{\infty} = r_{tt}X_1 + (1 - r_{tt})M_1 \tag{5.19}$$

where $\overline{X}_{\infty}$ = estimated true score; r_{tt} = reliability coefficient of the test; X_1 = obtained score; and M_1 = mean of the test scores.

The standard error of the estimated true score is given by the following formula:

$$\mathrm{SE}_{\infty} = \sigma\sqrt{r_{tt} - r_{tt}^2} \tag{5.20}$$

where SE_{∞} = standard error of the estimated true score; r_{tt} = reliability coefficient of the test.

To illustrate, suppose that Mohan has obtained a score of 50 on an intelligence test. The test has a reliability coefficient of 0.80 and the mean and the standard deviation of the test are 30 and 10 respectively. What would the true score of Mohan be? By Equation 5.18, the estimated true score will be $= (0.80)(50) + (1 - 0.80)(30) = 40 + (0.20)(30) = 40 + 6 = 46$. Thus Mohan's true score is 46 and his obtained score is 50. The standard error of this estimated true score by Equation 5.20 is:

$$\mathrm{SE}_{\infty} = 10\sqrt{0.80 - (0.80)^2} = 10\sqrt{(0.80 - 0.64)} = 4.$$

The 0.95 confidence interval (or 5% level of significance) for the true score would be $\pm\, 1.96 \times 4 = 7.84$ units above or below the estimated true score, that is, the true score will be between 53.84 and 38.16 (or 38–54 interval). Thus, only 5 scores in 100 will fall below 38.16 or above 53.84. Likewise, 0.99 confidence interval (or 1% level of significance) for the true score would be $\pm\, 2.58 \times 4 = 10.32$ units above or below the estimated true score, that is, the true score will be somewhere between 56.32 and 35.68 (or in 35–56 interval). Chances are that 1 score out of 100 will fall outside this range. In 0.95 confidence interval and 0.99 confidence interval, the range of the limits are 16 units and 21 units respectively. In the above example, although the reliability of the test is high (0.80), the estimation of true score has been subject to error. This indicates that the expert's attempt to estimate the true score of an examinee on the basis of the reliability coefficient of the test is not highly satisfactory.

INDEX OF RELIABILITY

In the previous section, the true score on a test was defined as the mean of an unlimited number of measurements made by the same person on the same test or on the equivalent forms of the same test. Theoretically, the correlation between the obtained scores and the true scores should be perfect. But in reality this is an exception rather than the rule. *Index of reliability* is statistically defined as the correlation coefficient between the obtained scores and their true counterparts. This statistic indicates the extent to which we can depend upon obtained scores as a measure of true scores. Index of reliability, thus, gives the maximum correlation which the test is capable of yielding in its present form. Index of reliability is statistically equal to the square root of the reliability coefficient of the test. Hence, its formula is

$$r_{1\infty} = \sqrt{r_{tt}} \qquad (5.21)$$

where $r_{1\infty}$ = the index of reliability; and r_{tt} = the reliability coefficient of the test.

To illustrate, suppose a test has a reliability coefficient of 0.65 and then its index of reliability through Equation 5.21 becomes $= \sqrt{0.65} = 0.806 = 0.81$, which means that this is the maximum correlation which this test can presently yield.

RELIABILITY OF DIFFERENCE SCORE

A researcher is often faced with a situation in which he has to look into differences in scores between two tests. For example, in evaluating a training program, he may be interested in knowing the difference between scores on a pretest and post test. Likewise, a counselling psychologist might be interested in knowing the difference between mathematics and general knowledge. In either case he would compute *change score* or *difference score* and he may be surprised to find that difference score shows low levels of reliability, even when the two tests in question are highly reliable. The reliability of such difference score can be estimated with the formula:

$$r_{DD} = \frac{\dfrac{r_{XX} + r_{YY}}{2} - r_{XY}}{1 - r_{XY}} \qquad (5.22)$$

where r_{DD} = reliability of the difference between scores on X and scores on Y

r_{XX} = reliability of X

r_{YY} = reliability of Y

r_{XY} = correlation between X and Y

A simple look at Formula 5.22 reveals a very important fact, that is, all other things being equal, *the higher the correlation between X and Y, the lower the reliability of the difference score.* To understand this apparent paradox, it is essential to know that the difference between X and Y reflects two important factors; namely, *differences in true scores* and *differences in measurement error*. If X and Y are highly correlated, the true-score part of X overlaps considerably with the true-score part of Y. Consequently, there will hardly be any difference in true scores on these two variables. Thus the differences between scores on X and Y will be due almost entirely to measurement error. Thus the more highly correlated X and Y are, the more reliable is their difference.

RELIABILITY OF COMPOSITE SCORE

A composite score, or summed score, is one which is obtained by adding the scores on several different tests. Researches have shown that whereas difference scores are often unreliable, composite scores are typically more reliable than the tests which make up the composite score . If individual tests are highly correlated, the reliability of the sum of those tests will be higher. The reliability of a composite score is estimated by Formula 5.23.

$$r_{ss} = \frac{k - (k_{\bar{r}_{ii}})}{1 - k + [(k^2 - k)_{\bar{r}_{ij}}]} \qquad (5.23)$$

where r_{ss} = reliability of composite score

k = number of test

$\bar{r}_{ii}$ = average test reliability

$\bar{r}_{ij}$ = average correlation between tests

The basic principle for determining the reliability of a composite score is very much similar to those involved in determining the internal consistency of a test. If more such tests are combined among which intercorrelation is high, they are likely to yield higher reliability of the composite score. In fact, adding together scores on several highly correlated tests achieves exactly the same goal as adding together scores on several positively correlated test items to form a single test score.

☑ Review Questions

1. Give the technical or logical meaning of reliability. Discuss any two methods of estimating reliability of test scores.
2. What is meant by internal consistency reliability? Discuss any two methods of assessing internal consistency reliability.
3. Discuss the relation between length of the test and reliability of the test. Is reliability also influenced by difficulty value of the item?
4. Discuss the factors that intrinsically affect the reliability of the test scores.
5. Discuss the major criteria of alternate-forms test. How can reliability of an alternate-forms test be increased?
6. Write short notes on the following:
   (a) Index of reliability
   (b) Homogeneity of test and its relation to reliability
   (c) K–R$_{20}$ and K–R$_{21}$
   (d) Rulon formula
   (e) Flanagan formula
   (e) Coefficient alpha
7. Outline the history and theory of reliability.
8. What is Bandwidth-Fidelity dilemma? How does it affect the reliability of the test?
9. Suggest measures for calculating reliability of difference score.
10. Outline a plan for calculating reliability of composite score.

◆

6

VALIDITY

CHAPTER PREVIEW

- Meaning of Validity
- Aspects of Validity
    - Content or Curricular Validity
    - Criterion-related Validity
    - Construct Validity
- Convergent Validation and Discriminant Validation
- Statistical Methods for Calculating Validity
    - Correlation Methods
    - Expectancy Tables
    - Cut-off Score
    - Miscellaneous Techniques
- Factors Influencing Validity
    - Length of the Test
    - Range of Ability (or Sample Heterogeneity)
    - Ambiguous Directions
    - Socio-cultural Differences
    - Addition of Inappropriate Items
- Concept of Cross-Validation
- Extravalidity Concerns
- Relation of Validity to Reliability

MEANING OF VALIDITY

Validity is another important characteristic of a scientific instrument. The term 'validity' means truth or fidelity. Thus, validity refers to the degree to which a test measures what it claims to measure. Validity is not the self-correlation of the test, rather it is the correlation with some outside independent criteria, which are regarded by experts as the best measure of the trait or ability being measured by the test.

Different writers have defined validity in different ways. For example, Anastasi (1968, 99) has said, "The validity of a test concerns what the test measures and how well it does so." Lindquist (1951, 213) has defined validity of a test as "the accuracy with which it measures that which is intended to measure or as the degree to which it approaches infallibility in measuring what it purports to measure." Kaplan and Saccuzzo (2001) have defined validity as "the agreement between a test score or measure and the quantity it is believed to measure." Paraphrasing the definition of validity from the influential *Standards for Educational and* *sychological Testing* (AERA, APA & NCME 1999) it can be said, "A test is valid to the extent that ferences made from it are appropriate, meaningful, and useful." These definitions point to the t that for determining the validity of a test, the test must be compared with some ideal ependent measures or criteria. The correlation coefficient computed between the test and the

ideal measures or criteria is known as the *validity coefficient*. 'Independent criteria' refer to some measure of the trait or the group of traits (outside the test) that the test itself claims to measure.

In a broad sense, validity is concerned with generalizability. When a test is a valid one, it means its conclusion can be generalized in relation to the general population. When the independent criterion has been set and when both the test and the criterion are reliable, the correlation between the test and the criterion can safely be taken as evidence of validity of the test. Although validity defined the meaning of tests and measures, the term itself started losing its meaning. However, in 1985, the joint committee of the American Educational Research Association (AERA), the American Psychological Association (APA) and the National Council on Measurement in Education (NCME) published a very important booklet for psychological test guidelines called *Standards for Educational and Psychological Testing* which was revised in 1999. This joint committee, by rejecting the numerous possible definitions of validity, suggested that validity is nothing but a simple evidence for inferences made about a test score. These evidences may be content-related, criterion-related or construct-related. In this way, validity really refers to evidence supporting what can be said on the basis of the test scores and not the tests themselves (Landy 1986). Validity has five important properties:

1. Validity is a relative term. A test is not generally valid. It is valid only for a particular purpose. For example, a test of statistical ability will be valid only for measuring statistical ability because it is put to use only for measuring that ability. It will be worthless for other uses like measuring the knowledge of geography, history, etc. It is obvious from this interpretation that strictly speaking, one validates not a measuring instrument, rather some uses to which the test is put.

2. Validity is not a fixed property of the test because validation is not a fixed process, rather an unending process. With the discovery of new concepts, and the formulation of new meanings the old contents of the test become less meaningful. Therefore they need to be modified radically in the light of the new meanings. Hence, the validity of a test computed in the beginning becomes less dependable and therefore, the test constructor should compute a fresh validity of the test in the light of the new meanings attached.

3. Validity, like reliability, is a matter of degree and not an all-or-none property. A test meant for measuring a particular trait or ability cannot be said to be either perfectly valid or not valid at all.

4. Validity is a unitary concept. In the two most important revisions done in 1999 of the *Standard of Educational and Psychological Testing* by the American Educational Research Association (AERA), the American Psychological Association (APA) and the National council on Measurement in Education (NCME), the view that there are different types of validity has been discarded. Instead, validity has been viewed as a unitary concept based on various kinds of evidences.

5. Validity involves an overall evaluative judgement. In other words, it requires an evaluation of the degree to which the interpretation and uses of test results are justified by supporting evidences as well as in terms of the consequences of those interpretations and uses. The *Standards for Educational and Psychological Testing* have provided five sources of evidences for evaluating the validity of a specific use or interpretation. These sources of evidences are (a) test content (b) response processes (c) internal structure (d) relations to other variables and (e) the consequences of testing. It means that the validity may include a consideration of the content measured, the ways of responding by the respondents, the relationship of the individual items to the test scores, relationship of the performance to other measures as well as the consequences of using the assessment.

ASPECTS OF VALIDITY

Since a test is valid to the extent that it serves the purpose for which it is to be used, and since there are many purposes of testing, it automatically follows that there are different aspects of validity representing each purpose of the test. Ordinarily, there are three main purposes of testing:

1. *Representation of a certain specified area of content*: The tester may wish to determine how an examinee performs at present in a sample of situations (or contents) that the test claims to represent. For example, through an English spelling test (a kind of achievement test) the tester may determine the present level of English spelling among school pupils.

2. *Establishment of a functional relationship with a variable available at present or in future*: The tester may wish to predict an examinee's future standing on a certain variable or he may wish to determine his present standing on a particular variable. For example, on a mechanical aptitude test he may wish to measure mechanical aptitude and predict his future performance in a job of a mechanic. Likewise, a tester may wish to determine the level of emotional adjustment of the examinee through the appropriate measure so that he may be able to infer his day-to-day adjustment with his peers.

3. *Measurement of a hypothetical trait or quality (or construct)*: A tester may wish to determine the extent to which an examinee possesses some trait(s) as measured by the test performance. For example, a tester may wish to know whether or not an examinee scores high on some abstract measures like extroversion, neuroticism, intelligence, which cannot be observed directly.

The above three purposes are basic to any testing programme. Corresponding to each of these purposes, there are three main types of validity. The typology is based upon the categories discussed in a monograph entitled *Standards for Educational and Psychological Tests and Manuals* published by the American Psychological Association (1966). One of the advantages of this typology is that different types of validity which generally create confusion, if treated separately, have been reduced to three main types. The three types of validity are: (i) content or curricular validity, (ii) criterion-related validity, and (iii) construct validity.

However, a word of caution has been sounded by the recent version of *Standards for Educational and Psychological Tests*. The consensus document of the *Standards* cautions against separation of validity into subcategories like predictive validity, criterion validity and content validity. It further opines that although categories for grouping different types of validity are convenient, the use of the categories does not mean that there are distinct *forms* of validity. Sometimes psychologists show overenthusiasm for making distinctions among categories when indeed the categories tend to overlap (Anastasi 1995; Messick 1998). The 1999 edition of *Standards* no longer recognizes the different categories of validity, rather it recognizes different categories of evidence for validity.

Content or Curricular Validity

Content validity is also designated by other terms such as *intrinsic validity, relevance, circular validity* and *representativeness*. Content validity is a nonstatistical type of validity that is usually associated with achievement tests. When a test is constructed so that its content of term measures what the whole test claims to measure, the test is said to have content or curricular validity. Thus content validity is concerned with the relevance of the contents of the items, individually and as a whole. Each individual item or content of the test should correctly and adequately sample or measure the trait or the variable in question and the test, as a whole, should contain only the representative items of the variable to be measured by the test. Anastasi (1968,100) has said that content validity, "involves essentially the systematic examination of the test content to determine whether it covers a representative sample of the behaviour domain to be measured." Content validity is needed in the tests which are contructed to measure how well the examinee has mastered the specific skills or a certain course of study.

In fact, content validity is the degree to which a test measures an intended content area. Psychometricians are of the view that content validity requires both *item validity* and *sampling validity*. Item validity is basically concerned with whether the test items represent measurement in the intended content area, and sampling validity is concerned with the extent to which the test samples the total content area. For example, a test designed to measure knowledge of biology might have good item validity because all the items indeed deal with good biology facts but might have poor sampling validity, that is, all its items may deal *only* with vertebrates. Thus a test with good content validity also samples the appropriate content area. This becomes important because we can't possibly measure each and every aspect of a certain content area. Yet the investigator wants to make inferences about performances in the whole content area based upon the performance of items included in the test. Such inferences can be correctly done only when test items adequately sample the domain of possible items.

Recently, two new concepts relevant to content validity have been introduced in the latest version of Standards for Educational and Psychological tests (AERA, APA & NCME 1999). These two concepts are: *construct under-representation* and *construct-irrelevant variance*. Construct under-representation means failure to incorporate important components of construct. For example, if a test of knowledge of history included only ancient history but not the modern one, the validity of the test would be threatened by construct under-representation. Construct-irrelevant variance is said to occur when scores are influenced by factors irrelevant to construct. For example, a test of intelligence might be influenced by test anxiety, reading comprehension or illness, etc.

Content validity of a test is examined in two ways: (i) by the expert's judgement, and (ii) by statistical analysis. Suppose the investigator wants to examine the content validity of a test on Indian history. For this purpose, the contents (or items) of the test will be submitted to a group of subject-matter experts. These experts will judge whether or not the items represent all the important events of Indian history, whether or not some additional items should be added for complete coverage, what should the relative weights of the items of a particular event be, etc. The validity of the contents or items will be dependent upon a consensus judgement of the majority of the subject-matter experts. Statistical methods may also be applied to ensure that all items measure the same thing, that is, a statistical test of internal consistency may provide evidence for the content validity. Another statistical technique for ensuring content validity may be to correlate the scores on the two independent tests, both of which are said to measure the same thing. Suppose one wants to know the content validity of a Hindi spelling test. Then the teacher can correlate the scores on the said test with another similar Hindi spelling test. A high correlation coefficient would provide an index for the content validity. Although a high correlation coefficient can easily be demonstrated in two sets of scores obtained from two similar tests, it does not fully guarantee content validity because high correlation may be due to the fact that both the tests measure the same incorrect things. Data relating to the item-discriminating power may also provide circumstantial evidence for the content validity. Items showing such power, that is, items discriminating among superior and inferior examinees are said to have content validity. The following points should be fully covered for ensuring full content validation of a test:

1. The area of content (or items) should be specified explicitly so that all major portions in equal proportion be adequately covered by the items. This specification should be followed rigidly for removing the general tendency of the item writers to include such items which are readily available and are easily written.
2. Before the item writing starts, the content area should be fully defined in clear words and must include the objectives, the factual knowledge and the application of principles and not just the subject matter.
3. The relevance of contents or items should be established in the light of the examinee's responses to those contents and not in the light of apparent relevance of the contents themselves. This is because the contents may appear to be relevant for a specific skill or

a certain course of study but may not be equally relevant to the examinees and then, they may misunderstand and give inappropriate responses. Thus, content validity is dependent not upon the apparent relevance of the test items, rather upon the relevance of the responses given by the examinees towards those test items.

Content validity is most appropriately applied to the achievement test or the proficiency test. For the aptitude test, the intelligence test and the personality test, content validity is not essential and sometimes may be a misleading index because the contents of these tests have less intrinsic resemblance or similarity to the trait or behaviour they are attempting to sample than do the achievement tests.

Face validity is often confused with content validity, but in the strict sense it is quite different. Face validity refers not to what the test actually claims to measure but to what it appears to measure superficially. In other words, face validity is the mere appearance that the test has validity (Kaplan & Sacuzzo 2001). Thus face validity should not be taken in the technical sense, nor should it be regarded as a substitute for objectively determined validity. When a test item looks valid to the group of examinees because it provides a logical link with the objective, the test is said to have face validity. Thus face validity is, in fact, a matter of social acceptability and not a technical form of validity like content validity, criterion-related validity and construct validity. Obviously, then the purpose of face validity is to establish rapport and secure co-operation because when test items do not appear to be valid to the examinees, they may not co-operate in responding and may give irrelevant answers because such items themselves appear to be irrelevant. It also keeps the examinees motivated. As Sackett et al. (2001) have rightly commented, "When a test looks appropriate for the performance situation in which the examinees will be expected to perform, they tend to react positively."

The above interpretation should not be taken to mean that the face validity is a useless quality. Face validity is needed in all types of tests and helps a lot in improving the objectively determined validity of the test by improving the wording and structure of the test contents.

Criterion-related Validity

Criterion-related validity is a very common and popular type of test validity. As its name implies, criterion-related validity is one which is obtained by comparing (or correlating) the test scores with scores obtained on a criterion available at present or to be available in the future. The criterion is defined as an external and independent measure of essentially the same variable that the test claims to measure. In this sense, by way of defining the validity of a test, Cureton (1965) has said that the validity of a test is an estimate of the correlation coefficient between the test scores and the "true" (that is, perfectly reliable) criterion scores. There are two subtypes of criterion-related validity: (a) *predictive validity*, and (b) *concurrent validity*. A detailed discussion of these two subtypes is given below.

Predictive Validity

Predictive validity is also designated as *empirical validity* or *statistical validity*. As its name implies, in predictive validity a test is correlated against the criterion to be made available sometime in the future. In other words, test scores are obtained and then a time gap (period) of months or years is allowed to elapse, after which the criterion scores are obtained. Subsequently, the test scores and the criterion scores are correlated and the obtained correlation becomes the index of validity coefficient. Marshall & Hales (1972, 110) have said, "The predictive validity coefficient is a Pearson product-moment correlation between the scores on the test and an appropriate criterion, where the criterion measure is obtained after the desired lapse of time." Let us take an example to illustrate predictive validity. Suppose one wants to predict success in a postgraduate class in terms of grade A, B, C, D and E—A being the best grade and E being the worst grade. The investigator may administer a test of intelligence at the time of their admission to

the postgraduate class and thus obtain a set of scores. After two years on the basis of classroom performance, students are graded according to the above categories. Here, grade points would constitute the criterion. A PM correlation through a scatter diagram may be computed between the set of intelligence scores and the grade point obtained after two years. If the correlation is high, we can say with all certainty that scores on intelligence are directly predicting the future performance of the students in the postgraduate class. The correlation becomes the index of the validity coefficient. Likewise, in an industry, the management may wish to select such workmen who can exhibit best performance on the job. For this purpose, they select a test which has high predictive validity. They may, for example, administer a test of special ability (or aptitude); say the mechanical aptitude test, for the selection of workmen. The scores on the mechanical aptitude test may be correlated after a year or two, against their average performance as measured in terms of units of production. If the correlation is high, the mechanical aptitude test is said to have predictive validity. Predictive validity is needed for tests which include long-range forecast of academic achievement, forecast of vocational success and forecast of reaction to therapy.

Concurrent Validity

Concurrent validity is another subtype of criterion-related validity. Concurrent validity is very similar to predictive validity, except that there is no time gap in obtaining test scores and criterion scores. The test is correlated with a criterion which is available at the present time. Scores on a newly constructed intelligence test may be correlated with scores obtained on an already standardized test of intelligence. The resulting coefficient of correlation will be an indicator of concurrent validity. If the correlation is too high, it will indicate that the new test is a needless duplication of the previous one. Likewise, an intelligence test may be validated (or correlated) against the marks obtained in the previous examination. This will also be an example of concurrent validity. Concurrent validity is most suitable to tests meant for diagnosis of the present status rather than for prediction of future outcomes.

Concurrent validity can be determined by establishing *relationship* or discrimination. The relationship method is simple and it involves determination of the relationship between scores on the test and scores on some other established criterion which are concurrently available. In this method, the steps involved are as follows:

(a) The test is administered to a defined group of individuals.

(b) The criterion or previously established valid test is also administered to the same group of individuals.

(c) Subsequently, the two sets of scores are correlated.

(d) The resulting coefficient indicates the concurrent validity of the test. If the coefficient is high, the test has good concurrent validity.

The discrimination method of computing concurrent validity involves determining whether the test scores can be used to discriminate between persons who possess a certain characteristic and those who don't possess such a characteristic, or between those who possess more of a certain characteristic and those who possess less of that characteristic. For example, a mental adjustment inventory would have concurrent validity if scores obtained on it could be used to discriminate correctly between institutionalized and noninstitutionalized persons.

A comparative study of predictive validity and concurrent validity has revealed that for the same test, predictive validity is usually lower than concurrent validity. The reason is that the degree of association between the test and the criterion *decreases* over time. The academic aptitude test scores will be more highly related with the present academic achievement of the examinees than with the academic achievement to be obtained five years later. Naturally, then predictive validity will be somewhat lower than concurrent validity. If concurrent validity of a test happens to be zero then its predictive validity is most likely to be zero or close to it.

Whether the test constructor is computing predictive validity or concurrent validity, he is faced with one vital problem. This problem, which is comparatively more acute in predictive validity, is the identification and selection of an appropriate and adequate criterion. An inappropriate and inadequate criterion may bring down the coefficient of correlation and thus, the validity of the test is adversely affected. For example, when one is validating a test of intelligence against a future criterion of grade (predictive validity), he may be faced with such a problem because something more than intelligence may be involved in obtaining a particular grade. Factors like interest, motivation, health, emotional adjustment of the students as well as the grading principles and practices, may influence the grades which are obtained by a student. In such situations, correlation between intelligence test score and the grade will not be a true index of validity of the test. Moreover, there are students who have low abilities but high achievements and high abilities but low achievements. In validating an intelligence test against the grades, such students are likely to bring down the coefficient of validity. Likewise, in the industrial situation, a test meant for selection of workers is usually validated against the criterion of the unit of production over a specified period of time. Again, we are faced with similar problems because units of production over a specified period of time may not be solely related to workers' ability, rather to a variety of factors like interest, motivation of the workers, adequate light and heat, speed of machines, regular supply of electricity and raw materials, etc. Likewise, when we are computing the concurrent validity of a test, the criterion may not be itself reliable to the extent it should. Again, in such a situation the correlation coefficient would tend to be attenuated or reduced and thus, the validity of the test would be substantially lower than the true or the theoretical relationship between the test and the criterion. It is, therefore, suggested that the obtained validity coefficient should be corrected for attenuation.

Correction for Attenuation

In computing criterion-related validity coefficients, two types of correction for attenuation are put into order: a full correction which includes correction in both the test as well as the criterion and one-way correction which includes correction in the criterion only. Sometimes both the test (which is being validated) and the criterion (against which the test is being correlated) are unreliable due to the faulty methods of test construction. Hence, the correlation coefficient between such a test and the criterion is automatically lowered. Therefore, a general formula which takes into account the unreliability (or measurement errors or chance errors) in test as well as in the criterion, is applied:

$$r_c = \frac{r_{xy}}{\sqrt{(r_{ttx})(r_{tty})}} \tag{6.1}$$

where r_c = correlation between test x and criterion y corrected for attenuation; r_{xy} = obtained validity coefficients, that is, obtained correlation between test x and criterion y; r_{ttx} = reliability coefficient of test x and, r_{tty} = reliability coefficient of the criterion y.

Equation 6.1 provides a full correction for those errors of measurement which tend to lower the reliability coefficients of both the test and the criterion. To illustrate, suppose the reliability coefficient of a test is 0.87 and that of the criterion is 0.80 and the obtained validity coefficient between the test and the criterion is 0.50. Now, the corrected validity coefficient, according to Equation 6.1, would be:

$$r_c = \frac{0.50}{\sqrt{(0.87)(0.80)}} = \frac{0.50}{\sqrt{0.696}} = \frac{0.50}{0.834} = 0.60$$

It is obvious from Equation 6.1 that the correction for attenuation increases the validity coefficient of the test. If the reliability coefficient of the test as well as criterion is 1.00, correction for attenuation in terms of Equation 6.1 has no meaning. Likewise, when corrected validity coefficient (that is, r_c) is either 1.00 or very close to it, students should note that the test and the criterion, then, are no longer different, rather, the two forms of the same test. Students should also note that the denominator in Equation 6.1 sets the upper limit of the correlation between the test and the criterion. In the above example, the denominator is equal to $\sqrt{(0.87)(0.80)} = \sqrt{0.696} = 0.834$, which indicates that in that particular example, the test and the criterion cannot correlate beyond 0.834.

It often happens that the criterion is far less reliable than the test itself. Hence the majority of experts are of opinion that the correction for attenuation should be applied only to the criterion and not to both. This is known as *one-way correction for attentuation*. When the correction for attenuation is to be applied only in the criterion, Equation 6.1 is reduced to:

$$r_c = \frac{r_{xy}}{\sqrt{r_{tty}}} \tag{6.2}$$

where symbols are defined as those for Equation 6.1. To illustrate, suppose the validity coefficient of a test is 0.45 and the reliability coefficient of the criterion is 0.67, then the corrected validity coefficient according to Equation 6.2, would be:

$$r_c = \frac{0.45}{\sqrt{0.67}} = \frac{0.45}{0.818} = 0.550 = 0.55$$

It is obvious, thus, that when correction for attenuation is applied to the criterion only, it tends to increase the validity coefficient of the test.

Major Qualities Desired in a Criterion Measure

In computation of validity coefficient, identifying a satisfactory criterion measure is a very important task. But this task is not so easy. For a satisfactory criterion measure, tests-and-measurements research workers have identified four desired qualities: *relevance, freedom from bias, reliability* and *availability*. A discussion follows.

(i) *Relevance:* A criterion with which the test is to be correlated must be relevant one. A relevant criterion very closely corresponds to the behaviours of the ultimate interest. We judge any criterion to be relevant to the extent that the standing on the criterion measure corresponds to the status on the trait one is going to predict. In general, relevance corresponds to the content validity of the criterion measure. For example, for a test of abstract reasoning, a criterion which closely assesses only abstract reasoning, will be considered as a relevant criterion.

(ii) *Freedom from Bias:* Another important quality desired in the criterion is the freedom from bias, which means that the measure should be such that each person has almost the same opportunity to score high or low, or should be such as that each person with equal capability obtains the same score irrespective of the group to which he or she belongs. A criterion measure that does not allow or restricts such opportunity or produces variations in scores among persons having equal capability, is said to be a biased one. Any such criterion that contains substantial bias cannot reveal relevant differences among people on the trait of interest (Comilli 2006). To the extent that criterion score depends on the personal characteristics of the person or on factors in the conditions of work rather than on the trait of interest, there is no meaning of the correlation between test scores and the scores on the criterion measure.

(iii) *Reliability:* Reliability of the criterion scores is another major quality. By reliability of the criterion scores is meant that the criterion score must not fluctuate from one situation to another, rather there must be stability or reproducibility in them. If the criterion score is such that it jumps around in an unpredictable way from one situation to another or from day to day so that a person who gets high score on one day or in one situation, gets low score on the other day or in other situation, there is no possibility that it will predict the score and yield a high correlation with the test.

(iv) *Availability:* Another desired quality of the criterion measure is that it should be conveniently available to the researcher who is determining the validity coefficient of the test scores. Criterion measure should not be such that one is required to wait for longer period of time nor should it be a very costly one. In either case, the criterion will not be easily available. Therefore, any choice of criterion must also take this practical limit into view.

Construct Validity

Construct validity is the third important type of validity. The term "construct validity" was first introduced in 1954 in the Technical Recommendations of the American Psychological Association and since then it has been frequently used by measurement theorists.

Construct validation is a more complex and difficult process than content validation and criterion-related validation. Hence an investigator decides to compute construct validity only when he is fully satisfied that neither any valid and reliable criterion is available to him nor any universe of content entirely satisfactory and adequate to define the quality of the test is present. In other words, construct validity is computed only when the scope for investigating criterion-related validity or content validity is bleak. Thus the process of construct validity is required when, "no criterion or universe of content is accepted as entirely adequate to define the quality to be measured, " (Cronbach & Meehl 1955, 282).

Construct validity has also been given other names such as *factorial validity* and *trait validity*. In construct validity, the meaning of the test is examined in terms of a construct. Anastasi (1968, 114) has defined it as "the extent to which the test may be said to measure a theoretical construct or trait." What is a construct? A construct is a nonobservable trait, such as intelligence, which explains our behaviour. According to Nunnally (1970), a construct indicates a hypothesis which tells us that "a variety of behaviours will correlate with one another in studies of individual differences and/or will be similarly affected by experimental treatments." A few examples of construct are anxiety, intelligence, verbal fluency, extroversion, neuroticism, dominance, etc. The process of validation involves the following steps:

I. Specifying the possible different measures of the construct
II. Determining the extent of correlation between all or some of those measures of construct
III. Determining whether or not all or some measures act as if they were measuring the construct

A brief discussion of each of these steps for determining construct validity follows:

I. *Specifying the possible different measures of the construct:* This is the first step in any construct validational study. Here the investigator explicitly defines the construct in clear words and also states one or many supposed measures of that construct. There is no standard way of stating the different measures of the construct. Specification of such measures is partly dependent upon the previous researches conducted in that area and partly upon the intution of the investigator. Suppose one wants to specify the different measures of the construct 'intelligence'. The investigator would have, first, to define the term 'intelligence' and in the light of the definition he would be expected to specify the different measures. He may define intelligence as an overall ability to think in a rational way, to act with a purpose and to make adjustment according to the varying demands of the environment. From this definition, a number of specifications may be made.

1. Quick decisions even in complex and difficult tasks
2. Ability to operate with symbols, figures, etc.
3. Ability to learn
4. Purposeful behaviour, or goal-related behaviour
5. Comprehension of verbal materials

6. Capacity to modify behaviour
7. Verbal fluency
8. Original and critical thinking
9. Ability to do numerical calculations rapidly

The above list of measures is not an exhaustive one. Other measures may be included. One general demerit is that if different investigators outline the number of measures of the same construct, they may not be in complete agreement among themselves regarding the specification of measures. But this demerit applies only to those constructs which have not been well defined and where important facts have not been unearthed.

II. *Determining the extent of correlation between all or some of the measures of construct:* When adequate measures of the construct have been outlined, the second step consists of determining whether or not those well-specified measures actually lead to the measurement of the concerned construct. This is done through an empirical investigation in which the extent to which the various measures 'go together', or correlate with each other, is determined. If the various measures highly correlate with each other or if the various measures are affected by a variety of experimental manipulations (called independent variables) in much the same way, we get sufficient evidence for concluding that they are measuring the same thing (that is, the same construct). In an empirical investigation as above (where the correlation coefficients are computed between the different measures of a construct), we may have to face any of the following facts:

(a) Some measures may correlate highly with some measures, but correlate very little with other measures present there. In other words, some measures tend to form a cluster and all members of a cluster correlate highly with each other and poorly with the members of other clusters. For example, out of the above nine measures the three measures, that is, 2, 5 and 9 may tend to form a cluster because these three are expected to correlate highly with each other but not to the same extent with other measures. Likewise, measures 1, 3, 4 and 6 may form one cluster and measures 5, 7 and 8 may tend to form another cluster.

(b) It may also be that all measures of a construct correlate highly with each other. This is an ideal situation for investigating construct validity because in such a situation, the investigator can safely conclude that all the proposed measures directly correlate with the construct.

(c) The possibility that all the proposed measures correlate very little or all correlations are near zero may also exist. If such is the result, the obvious conclusion is that the measures are all independent and each of them measures a separate ability or trait.

In practice, however, such a conclusion is rarely reached and the investigator is often thrown into a dispute before a final decision is taken.

III. *Determining whether or not all or some measures act as if they were measuring the construct:* When it has been determined that all or some measures or referents of construct correlate highly with each other (providing sufficient evidence for the referents that they all measure the same thing), the next step is to determine whether or not such measures behave with reference to other variables of interest in an expected manner. If they behave in an expected manner, it means they are providing evidence for the construct validity. For example, highly correlating measures from among the above nine supposed referents for intelligence should show at least some moderate correlation with teachers' ratings, grades in class and examination marks.

It is obvious from the above interpretation that unlike content validity and criterion-related validity, the evidence for construct validity is always circumstantial rather than direct. Construct validation is also a difficult process because it contains several problems like systematic examination concerning the definition of the construct, unsuitability and inappropriateness of the measures of the construct, lack of high correlations among measures, etc. This is the reason why construct validity is computed by the test constructor only if there is no suitable alternative.

There is no single method for evaluating construct validity. However, the evidence for this type of validity always depends upon a programme of research. Gregory (2005) has provided some examples of the kinds of findings that would indicate that a new test has construct validity.

(i) The test appears to be homogeneous and therefore, measures a single construct.
(ii) The test correlates highly with related tests/instruments/variables than with unrelated tests/instruments/variables.
(iii) Development changes over time or across periods of the different ages are consistent with the theory of construct being assessed.
(iv) Differences among the well-defined groups on the test are theory-consistent.
(v) Intervention effects produce changes on the test scores that are theory-consistent.
(vi) The factor analysis of the test scores produces results that are understandable in the light of the theory by which the test was constructed.

As an overall conclusion, it can be said that construct-related validity evidence, in reality, subsumes all other types of validity evidences. For example, in construct validation, content-related validation is an essential step. Likewise, for providing construct-related evidence for validity requires validation against many criteria. For example, a test of health status might be validated by establishing correlation with symptoms, doctor visits or other physiological variables. Many textbooks, until very recently, tried to divide validity into different types. But the reality is that any attempt to divide validity into categories like construct- and criterion-related validity is simply confusing because there is much similarity between construct validity and criterion-related validity. Today many psychologists believe that construct-related evidence for validity is the only real type of validity with which we should show concern because this validity subsumes the other so-called type of validity. According to the famous testing pioneer, Lee Cronbach, it is not appropriate to divide validity into three parts. Cronbach (1989, 91) has rightly maintained," All validation is one and in a sense, all is construct validation." The 1999 edition of *Standards for Educational and Psychological Testing* also rejects the different categories of validity and instead recognizes different categories of evidence for validity.

CONVERGENT VALIDATION AND DISCRIMINANT VALIDATION

Campbell & Fiske (1959) have demonstrated that the convergent validation and the discriminant validation are important for establishing satisfactory construct validity. For a satisfactory construct validity, it is essential that the test correlates with other theoretical measures with which it should correlate and it is equally essential that it must not correlate with other measures with which it should not. For example, a numerical aptitude test should correlate with an arithmetical reasoning test and algebra test but it should not correlate with a history or geography test. When the test correlates with its expected referents, the process is known as *convergent validity* and when a test correlates poorly (or not at all) with measures with which it should not, because it differs from those referents or measures, the procedure is called *discriminant validation*. In this sense, the low correlation sometimes becomes the evidence for the validity of the test.

In this way, it can be said that when a measure correlates well with other tests believed to assess the same construct, convergent validity is obtained. The evidence for convergent validity is obtained in one of two ways. First, we demonstrate that a test measures the same things as other tests used for the same purposes. Second, we demonstrate a specific relationship that we can expect that the test is really doing its job correctly. For demonstrating discriminant validity, the test should have low correlations with measures of unrelated constructs or evidence for what the test does not measure.

Campbell & Fiske (1959) have proposed a special experimental design for studying convergent and discriminant validity. The design is referred to as the multitrait-multimethod matrix (MTMM). As its name implies, the use of this matrix requires that two or more than two traits be measured by two or more than two methods. In this matrix, all the intercorrelations resulting from assessment of each of the traits by each of the methods are given. The essence of MTMM for assessing construct validity is that the scores resulting from different ways of assessing the same trait should correlate more highly with each other than they do with scores for other traits by the same method. The following hypothetical examples (given on the pattern of Campbell & Fiske 1959) serve as illustrations.

In Table 6.1, three traits (A, B, and C) have been measured by three methods (1, 2 and 3). The three traits may be aggression (A), neuroticism (B), and emotional stability (C), and the three methods may be personality inventory (1), projective test (2), and the experts' ratings (3). Accordingly, A_1 indicates aggression scores on personality inventory; A_2 indicates aggression scores on projective test; A_3 indicates aggression scores on the experts' ratings. Likewise, C_1 indicates scores of emotional stability on personality inventory and C_3 indicates scores of emotional stability on the experts' ratings.

Table 6.1 A Hypothetical Multitrait-Multimethod Matrix

		Method 1			Method 2			Method 3		
	Traits	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
Method 1	A_1	(0.80)								
	B_1	0.52	(0.90)							
	C_1	0.39	0.37	(0.86)						
Method 2	A_2	***0.58***	0.23	0.09	(0.96)					
	B_2	0.23	***0.58***	0.11	0.65	(0.94)				
	C_2	0.12	0.12	***0.56***	0.59	0.55	(0.95)			
Method 3	A_3	***0.57***	0.22	0.12	***0.67***	0.42	0.36	(0.94)		
	B_3	0.23	***0.59***	0.12	0.44	***0.67***	0.34	0.67	(0.92)	
	C_3	0.12	0.11	***0.55***	0.34	0.32	***0.60***	0.55	0.60	(0.95)

In Table 6.1, all correlation values are hypothetical and include the three sets of reliability diagonals (or monotrait-monomethod values) given in parentheses and the three sets of validity diagonals (or monotrait-heteromethod values) given in italics. The correlation values between the different traits measured by the same method have been shown by solid-line triangles (that is, by heterotrait-monomethod triangles) and the correlation values between different traits measured by different methods have been shown by broken-line triangles (that is, by heterotrait-heteromethod triangles). The cells of the matrix that contain value of correlations in brackets are called as *reliability diagonals* and all the cells having correlations between the two broken traingles are called as *validity diagonals*. The reliability diagonals and the nearly heterotrait-monomethod triangles constitute monomethod block and the validity diagonals and the two heterotrait-heteromethod triangles, one on each side, constitute the heteromethod block. A careful examination of Table 6.1 reveals the following four points which are important for establishing convergent validity and the discriminant validity (thus, establishing the construct validity):

1. The values in the validity diagonal should be larger and statistically significant ones. This would be the evidence for the convergent validity.

2. The values in the validity diagonals should be higher than values in heterotrait-heteromethod triangles (shown by broken-line triangles). In other words, the validity coefficient should be higher than the correlation between traits measured by the different methods. This would be the evidence for discriminant validity and the data of Table 6.1 fully meet this requirement.

3. The values in the validity diagonals should be higher than the values in heterotrait-monomethod triangles shown by solid lines. In other words, the validity coefficient should be higher than the correlation between different traits measured by the same method. If the opposite is the case, it means that the examinee's scores on the concerned method (may be the inventory or the projective technique or the experts' ratings) are affected by some irrelevant common factors lying within the examinee such as his social desirability (or undesirability) tendency or his ability to comprehend the language of the items, etc. Data of Table 6.1 partially meet the above requirement. Fulfilment of this requirement provides evidence for discriminant validity.

4. Both the types of heterotrait triangles, that is, heterotrait-monomethod triangles and heterotrait-heteromethod triangles should show within themselves some marked pattern of trait intercorrelation. This would also provide evidence for discriminant validity. Data of Table 6.1 meet this requirement to a greater extent.

If the validity coefficients meet the above requirements, they are said to have convergent validity and discriminant validity, which ultimately guarantee the establishment of construct validity. For a satisfactory construct validity, the requirements shown by at least the first three points must be met.

The Unified View of Validity

As discussed earlier, a new unified view of validity gradually emerged in the 1980s. According to this unified view, test validity is something of a misnomer because what is validated are the *interpretation* as well as the uses of test scores for a particular applied purpose, not the test itself (Cronbach 1988; Messick 1989; Shepard 1993). Validity requires an evaluation of the extent to which interpretations and uses of test results are justified by supporting evidence and in terms of the consequences of those interpretation and uses. Therefore, it involves an overall evaluative judgement (Linn & Miller 2011).

If the test constructor constructs a test and never uses it, validation is not an issue. If he administers the test to a group of examinees and obtains scores but locks those scores in an almirah, again validation is not an issue. But suppose he takes those test scores and uses them for a purpose, one may ask about or infer about the appropriateness of the test scores for that purpose or use.

Inferences related to the need of validation are of two types: *interpretative inference* and *action inference*. By interpretative inference is meant what the test scores mean, that is, how do we interpret them. By action inference is meant the claim regarding the appropriateness and utility of test scores forming the basis of some specific action.

In view of what has been said above, test validation is understood as simply the process of gathering evidence in support of those score-based inferences. Validation of interpretative inferences aims at bringing together several evidences in support of certain interpretation of test scores as well as demonstrating that other interpretations are not much plausible. Validation of action inferences requires evidences regarding the meaning of test scores, their appropriateness and usefullness for particular applied purposes. Thus, the goal of test validation is to build and develop a strong case for the inferences the researcher would like to make.

Messick (1989) has expanded the boundaries of validity beyond meaning of test scores to include value implications of score interpretation and utility. He strongly emphasized that validation must go beyond evidence related to the meaning of test scores so that the consequences of score interpretation and test uses (both actual and potential) must also be covered. This is popularly known as *Messick's expanded theory of validity*. According to this theory, the concept of validity is expanded beyond issues of test score meaning to include evidences of relevance and utility of test scores as well as the consequences of test interpretation and uses.

STATISTICAL METHODS FOR CALCULATING VALIDITY

There are several statistical methods that are employed in computing the validity coefficient of a test. A few important ones are enumerated below.

Correlation Methods*

The validity of a test is defined as its correlation with some outside independent criterion. The correlation coefficient can be calculated by different methods, the important ones being Pearson *r*, biserial *r*, point biserial *r*, tetrachoric *r*, phi coefficient and the multiple correlation. Of these, the use of Pearson *r* is very common. The biserial *r* and the point biserial *r* are used where one of the variables being correlated is divided into "Pass-Fail", "True-False", "Above median-Below median", etc., and the other is given in terms of continuous scores. When, on the other hand, scores on the test as well as on the criterion are divided into two categories, the tetrachoric *r* or the phi coefficient are the most appropriate statistics. Multiple correlation is used where more than two measures are involved. *R* is a symbol for multiple correlation which indicates the relationship between one measure and the composite of two or more than two sets of measures.

Expectancy Tables

The expectancy table is one way of showing the relation between the test scores and the criterion measures. In an expectancy table, the expectancy (usually expressed in terms of percentage or proportion) of the criterion measures for each examinee is given against each test score. Thus a sort of bivariate distribution is prepared in which test scores are plotted against the criterion scores or measures. From this bivariate distribution, the correlation coefficient (that is, validity coefficient) is computed. As its name implies, through the expectancy tables the predictive efficiency of a test is estimated. Estimates are usually based upon the probability that an examinee securing a particular score on the test will obtain a specified score or rating in the performance being predicted. Expectancy tables, therefore, answer questions like: What is the probability that a student with an IQ of 100 will secure the first rank in the annual examination of class X? What is the probability that a person securing the 90th percentile in the clerical speed and accuracy test would achieve a rating of "excellent" when posted as clerk? What is the probability that a person securing a score of 60 (out of a maximum of 139) in the medical entrance test would be rated as "excellent" after the completion of the course? Expectancy tables, intended to answer all such questions, would not be uniform. The form of the data in the tables would depend upon the particular probabilities that one wants to determine. However, al expectancy tables would have one thing in common, that is, the estimates of the probabilities o the score or rating on performance is predicted when the actual score is known. To illustrate th use of expectancy tables, an example with hypothetical data has been provided in Table 6.2.

---

* For a detailed discussion of the different correlational methods, students are advised to consult Guilford (1965) or McNem (1962).

Table 6.2 Expectancy table showing relationship between medical entrance test and criterion grade of 200 students of Patna Medical College Hospital

MET Scores	Number of students ($N = 200$)	Percentage in each criterion grade				
		Worse	Poor	Average	Good	Excellent
120–139	10				20	80
100–119	20			10	20	12
80–99	35			40	50	8
60–79	80	5	20	20	7	
40–59	25	10	30	15	3	
20–39	18	25	40	10		
0–19	12	60	10	5		
		100	100	100	100	100

It is obvious from the expectancy table in Table 6.2 that of the 10 students securing scores in between 120–139 in the Medical Entrance Test (MET) , 80% were rated as "excellent" and 20% were rated as "good". Likewise, of the 20 students securing scores in between 100–119 on the MET, 12% were rated as "excellent", 20% were rated as "good" and 10% were rated as "average". Similarly, on the lower extreme, there were 12 students out of whom 60% were rated as "worse", 10% were rated as "poor" and 5% were rated as average.

Cut-off Score

Cut-off score is defined as that score on the test which separates or demarcates between the potentially superior and the potentially inferior examinees. The purpose of using a cut-off score is to separate the maximum number of potentially superior or good examinees and to eliminate the maximum number of inferior or poor examinees. The general demerit of using a cut-off score is that the screening by means of such a score may not be perfect, partly because no test is perfectly valid, and partly because the actual ability or potential of the examinees may not be revealed in a single test. As a consequence, some inferior examinees may be selected whereas some superior examinees may be rejected.

Miscellaneous Techniques

Besides the above techniques for establishing the validity of the test, there are some additional procedures which are infrequently utilised. The comparison of test scores of the contrasted groups (such as normals and neurotics, institutionalized mental defectives and normal school-going children) is one way of establishing validity. Likewise, the percentage of the students successful in a particular age group may be grouped with the similar percentage in another age group for establishing the validity of the test.

FACTORS INFLUENCING VALIDITY

Validity of a test is influenced by several factors. Some of the important factors are enumerated below.

Length of the Test

Homogeneous lengthening of the test not only increases the reliability but also the validity of the test. The longer the test, the more reliable and valid it becomes. Thus, lengthening the test or repeated administration of the same test increases the reliability, and since validity in a homogeneous test is dependent upon reliability, lengthening also increases the validity of the test. But validity as compared to reliability does not change rapidly with increase in the length of the test or several repeated administrations of the same test. The increased validity of the test after homogeneous lengthening is given by the following formula:

$$r_{c(nx)} = \frac{(n)(r_{cx})}{\sqrt{n + n(n-1)r_{11}}} \tag{6.3}$$

where $r_{c(nx)}$ = correlation between criterion and the test lengthened n times; r_{cx} = correlation between the criterion and the test in original length; n = number of times the test has been lengthened or repeated on the same sample; and r_{11} = reliability coefficient of the test.

To illustrate, suppose a test has a validity coefficient 0.50, a reliability 0.40 and it is lengthened four times its present length. Then its validity, according to Equation 6.3, would be:

$$r_{c(nx)} = \frac{(4)(0.50)}{\sqrt{4 + 4(4-1)0.40}} = \frac{2}{\sqrt{4 + 4(3)(0.4)}}$$

$$= \frac{2}{\sqrt{(8)(3)(0.4)}} = \frac{2}{\sqrt{9.6}} = \frac{2}{3.098} = 0.64557 = 0.65$$

Thus, lengthening the present test four times or averaging four repeated administrations of the same test increases the validity coefficient from 0.50 to 0.65.

Sometimes, it becomes necessary to know how many times the test should be lengthened in order to reach a certain specified level of validity coefficient. In such a situation, Equation 6.3 may be rearranged as shown in Equation 6.4 so that it may be known beforehand how many times the test should be lengthened.

$$n = \frac{r^2_{c(nx)}(1 - r_{11})}{r^2_{cx} - r^2_{c(nx)}(r_{11})} \tag{6.4}$$

where n = number of times the test should be lengthened or repeated on the same sample; $r_{c(nx)}$ = desired level of validity coefficient; r_{cx} = validity coefficient between the criterion and the test; and r_{11} = reliability coefficient of the test.

To illustrate, suppose a test has a reliability coefficient of 0.40 and validity coefficient of 0.50. How many times should the test be lengthened in order to reach a validity coefficient of 0.65? In order to answer this, Equation 6.4 may be solved as follows:

$$n = \frac{(0.65)^2(1-0.40)}{(0.50)^2 - (0.65)^2(0.40)} = \frac{(0.42)(0.60)}{(0.25) - (0.42)(0.40)}$$

$$= \frac{0.252}{0.25 - 0.168} = \frac{0.252}{0.082} = 3.07 = 3$$

Thus, in order to reach the desired level of validity coefficient of 0.65, the present test should be lengthened three times.

Equation 6.3 gives an estimate of the validity coefficient only when the test has been lengthened n times. But sometimes the investigator increases the length of the test as well as the length of the criterion. In such a case, Equation 6.3 would not be a correct estimate of the validity coefficient, and for such a situation another formula is suggested:

$$r_{c(n_x m_c)} = \frac{r_{cx}}{\sqrt{\left(\frac{1-r_{11}}{n} + r_{11}\right)\left(\frac{1-r_{11c}}{m} + r_{11c}\right)}} \tag{6.5}$$

where $r_{c(n_x m_c)}$ = correlation between the test lengthened n times and criterion lengthened n times; r_{11} = reliability of the test; n = number of times the test has been lengthened; m = number of times the criterion has been lengthened; r_{1c} = reliability coefficient of the criterion; r_{cx} = correlation between the test and the criterion.

To illustrate, suppose a test has been lengthened three times and the criterion has also been lengthened three times, the reliability coefficient of the test and the criterion are 0.58 and 0.56 respectively, and the correlation coefficient of the test with the criterion was 0.65, its validity coefficient according to Equation 6.5 would be:

$$r_{c(n_x m_c)} = \frac{0.65}{\sqrt{\left(\frac{1-0.58}{3}+0.58\right)\left(\frac{1-0.56}{3}+0.56\right)}}$$

$$= \frac{0.65}{\sqrt{\left(\frac{0.42}{3}+0.58\right)\left(\frac{0.44}{3}+0.56\right)}} = 0.92$$

Thus, lengthening the test and the criterion three times the original length increases the validity coefficient from 0.65 to 0.92. Now one question arises: What would happen to the reliability and the validity of the test if the test is lengthened indefinitely? If the test is lengthened indefinitely, its reliability would be perfect but validity would not be perfect because the variance of the test may still have some distinctiveness or specificity. The maximum validity when the test is lengthened indefinitely can be estimated by the following formula:

$$r_{\infty xc} = \frac{r_{xc}}{\sqrt{r_{11}}} \tag{6.6}$$

where $r_{\infty xc}$ = maximum validity when the test has been lengthened indefinitely; r_{xc} = obtained correlation (or validity coefficient) between the test and the criterion; r_{11} = reliability coefficient of the test.

A close scrutiny of Equation 6.6 reveals that it estimates the maximum validity of the test (when the test has been lengthened indefinitely) through the index of reliability, which is the denominator of the formula.

Range of Ability (or Sample Heterogeneity)

Like reliability, validity is also influenced by the range of ability of the samples used. If the subjects have a very limited range of ability (that is, the wider range of scores is not possible), the validity coefficient will be low. On the other hand, if the subjects have a wider range of ability so that a wider range of scores is obtained, the validity coefficient of the test would be enhanced.

Ambiguous Directions

If the directions of the test are ambiguous, it would be differently interpreted by different examinees. Moreover, such items tend to encourage guessing on the part of the examinees. As a consequence, the validity of the tests would be lowered.

Socio-cultural Differences

Cultural differences among different societies are likely to affect the validity of a test. A particular test developed in one culture may not be valid for another culture because of the differences in socio-economic status, sex ratios, social norms, etc. However, when a test is cross-cultured, this factor does not affect the validility of the test.

Addition of Inappropriate Items

When inappropriate items, particularly vague ones whose difficulty values differ widely from the original items, are added to the test, they are likely to lower both the reliability and the validity of the test.

CONCEPT OF CROSS-VALIDATION

Psychometricians are of the view that validation of a completed test should always be computed on a new sample of the same population, that is, upon a sample different from that used in item analysis. This process is called *cross-validation*. Thus the term cross-validation clearly refers to the practice of revalidating a test on a new sample of examinees. The process of cross-validation becomes necessary because the validity of the test when computed from the standardization sample will be spuriously high because items are so selected as to maximize differences between high and low scores. Apart from this, the validity coefficient of the test will be increased by chance factors peculiar to the standardization sample. In fact, a high validity coefficient might result under such circumstances even when the test has little validity. Thus, any validity coefficient computed on the same sample that was used in item analysis will capitalize on random sampling errors within that particular sample and will be spuriously high.

A classic demonstration of the use for cross-validation using empirical data has been provided by an early investigation conducted with the Rorschach test (Kurtz 1948). The Rorshach test was administered to 80 sales managers with a basic purpose of determining whether the test could be of any help in selecting such managers for life insurance agencies. In fact, these managers had been selected from several hundred employed by eight life insurance companies so as to represent an upper criterion group consisting of 42 managers considered very satisfactory by their companies and a lower criterion group of 38, considered unsatisfactory by their companies. All the eighty test records were studied by the Rorschach experts who selected a total of 32 signs or response characteristics, occurring more frequently in one criterion group than in the other. Signs found more often in the upper criterion group were given a score of +1 if present and 0 if absent. Similarly, those more common in the lower group were given −1 if present and 0 if absent. When the scoring key based on these 32 signs was reapplied to the original group of 80 persons, 79 of 80 managers were classified rightly as being in an upper or lower group. Thus the correlation between the test score and the criterion was nearly perfect, that is, 1.00. However, when the test was administered on a new sample of 41 managers (21 in the upper and 20 in the lower group) for the purpose of cross-validation, the validity coefficient dropped to .02. Thus, it became apparent that the scoring key developed in the first sample had no validity for selecting such personnel.

The effect of chance factors upon validity that may operate due to the use of a single sample for item selection and test validation can be shown through this example: suppose that the items of an aptitude test, which intend to select retail salesmen, have been so selected as to yield satisfactory validity indices in terms of the top 27% and bottom 27% of a sample of sales personnel. In fact, some irrelevant factors may be presented in this group and some of which may be correlated with scores on the test. Such factors will often be correlated with responses to the items in one extreme group more than in the other group. If the test is validated using the same sample, it is very likely that those incidental factors will make the validity coefficient spuriously high. But if the test is cross-validated on a new group of retail salesmen and unless those incidental factors are present in this new group to the same degree (which is not likely), the validity coefficient of the completed test will be lower in the new groups than in the original standardization group. In this way validity coefficient tends always to be unnecessarily high in the standardization group, thus making cross-validation necessary. Cureton (1950) had earlier also established that the use of a single sample for item selection and test validation can produce a completely spurious validity coefficient under pure chance condition.

Now the question: What are those factors upon which the amount of shrinkage of a validity coefficient in cross-validation depends? As we know, by validity shrinkage is meant a phenomenon wherein a test predicts the relevant criterion less accurately with a new sample than with the original standardization sample. The answer to the question just cited is that such shrinkage depends partly upon the size of the original item pool as well as upon the proportion of items retained finally and partly upon the size of the sample. When the initial item pool is large and the proportion of items retained is small, there is greater opportunity to capitalize on chance differences and therefore, shrinkage of test validity in cross-validity will be the greatest. Likewise, when size of the samples is small, there will be the greater validity shrinkage during cross-validation. It has also been observed that when items are assembled without previously formulated rationale, the shrinkage in validity coefficient in cross-validation will be greater. Mitchell and Klimoski (1986) have neatly demonstrated the differences in validity shrinkage that really occur when items are chosen by rational or by empirical strategies.

EXTRAVALIDITY CONCERNS

Actual uses of psychological test have revealed that even if a test is fair, valid and unbiased, the decision to use it is governed by several considerations. These considerations or concerns are technically called *extravalidity concerns*. Cole and Moss (1989) have pointed out the following four major extravalidity concerns:

(i) What is the purpose for which the test is used?
(ii) To what extent are those purposes accomplished by the actions done?
(iii) What are the likely side effects or unintended consequences of using a test?
(iv) What possible alternatives to the test may serve the same purpose?

Of these four major types of extravalidity concerns, the side effects and unintentional consequences of testing are important and therefore, we shall concentrate upon it. Those readers who want to refer to other types should consult Cole and Moss (1988), Cronbach (1989) and Jensen (1980).

When a psychological test is used, of course it produces the intended effect. But apart from this, some unintentional side effects are produced. Such effects cost a lot and the examiner has to determine whether the benefits of giving the test outweigh the costs of potential side effects. The consideration of these side effects should influence the examiner's decision to use a particular test for a particular purpose. The examiner may be forced not to use a test for a worthy purpose if he feels that the likely costs from the unintended side effects outweigh the expected benefits.

The unintended side effects of testing may be illustrated through some examples. Cole and Moss (1989) have cited the example of using psychological tests for determining eligibility for special education. Although the intended purpose of using psychological tests here is to help students in learning and strengthening their mental status, such process may produce several types of unintended negative side effects:

(i) Teachers may not tolerate these children and view them as unworthy of attention.
(ii) The identified children may feel unusual.
(iii) Other Children may call such identified children names.
(iv) Teachers may have to arrange a special class segregated by race or social class.

Likewise, suppose that MMPI (Minnesota Multiphasic Personality Inventory) is used to screen candidates for the job of police officer. As we know, MMPI basically measures pathological traits or dimension of personality. Those candidates who become unsuccessful due to high score on some dimensions of MMPI, such as Psychopathic Deviate, Schizophrenia, Depression, Hypochondriasis, etc., may be tagged with a pathological label which may debar them from being appointed in other fields too. The repercussions are not only unfair to the candidates but they also raise the probability of lawsuits against the employees.

Due to renewed sensitivity to extravalidity concerns, many psychometricians have preferred to widen their definition of test validity. Instead of defining test validity in technical terms, that is, a test is valid if it measures what it intends to measure, they have preferred a functionalist definition of validity that states that a test is valid if it serves the purpose for which it is used (Cronbach 1988; Messick 1995). Such functionalist perspective recognizes that the test validator has an obligation not only to enforce constructive consequences but also to guard against adverse outcomes (Messick 1980).

Messick (1995) and other supporters of functionalist perspective have opined that this new wider concentration of validity rests on the following four bases:

(i) Sufficient evidence for construct validity, that is, for appropriate convergent validity and discriminant validity
(ii) Analysis of the value implications of the test interpretation
(iii) Evidence for the usefulness of test interpretation
(iv) Assessment of the potential and actual social consequences including negative side effects of the use of the test

According to functionalist perspective, a valid test is one that answers well to all these four aspects.

RELATION OF VALIDITY TO RELIABILITY

Reliability and validity are two dimensions of the same thing, that is, test efficiency. Validity is the correlation of the test with some outside independent criteria and reliability of the test is the self-correlation of the test. A test which is not correlating with itself is not expected to correlate with outside independent criteria. In other words, a test which has poor reliability is not expected to yield high validity. Thus, validity is dependent upon reliability. This prediction is true for the homogeneous test only. If a test is heterogeneous (such as biographical-data blank), validity may be high even without high reliability (particularly the internal consistency reliability). This is because in a heterogeneous test, each part measures an independent function. As such the self-correlation (or the reliability) of the test, particularly the internal consistency reliability, would be extremely low because items would not highly correlate with each other. But when correlated with other sets of independent items, they are expected to yield a high correlation. Thus, validity may be high without the underlying high reliability. Reliability is a sufficient but not a necessary condition for validity. The validity of the test may be higher than the reliability but not higher than the index of reliability because the index of reliability sets the maximum limit of correlation that the test can yield with its true measures—the true measures being the criterion. Theoretically, the maximum validity coefficient between two variables is equal to the square root of the product of their reliabilities, that is, $r_{12\max} = \sqrt{r_{11}r_{22}}$ where r_{11} and r_{22} are reliabilities of the two tests/variables. For example, if the reliability of test is 0.8 and the reliability of the criterion is 1.0, then the maximum validity would be $\sqrt{0.8 \times 1} = \sqrt{0.8} = 0.89$.

A test constructor should not *always* aim at having high reliability and high validity in the same test. If he does so, he is said to be working at cross purposes because sometimes the goals of reliability and validity are incompatible. This is because the requirements for high reliability and high validity are opposite to each other. High reliability requires items of equal difficulty and high intercorrelations between the items, whereas high validity requires items of different difficulty values and low intercorrelations among items (Guilford 1956). Obviously, attempting high validity as well as high reliability would imply working at cross purposes. Fortunately, Tucke (1946) has provided a solution to this puzzle. He has provided a compromising and satisfactor method. According to him, if inter-item correlations range from 0.10 to 0.60, one can expect have both the reliability and the validity to a satisfactory degree.

☑ Review Questions

1. Distinguish between reliability and validity of test scores. Discuss the significance of different types of criterion-related validity of scores.
2. What is meant by construct validity? Discuss the steps involved in establishing a construct validity.
3. Discuss the factors that influence the validity of test scores.
4. Write short notes on the following:
    (a) Content validity
    (b) Convergent validation and Discriminant validation
    (c) Correction for attenuation
    (d) Relation of validity to reliability
5. Define the concept of cross-validation. Discuss its relevance for test construction.

◆

7

NORMS AND TEST SCALES

CHAPTER PREVIEW

- Meaning of Norm-Referencing and Criterion-Referencing
- Steps in Developing Norms
- Types of Norms and Test Scales
    - Age-equivalent Norms
    - Grade-equivalent Norms
    - Percentile Norms (or Percentile-rank Norms)
    - Standard Score Norms
- Computer Applications in Psychological Testing and Assessment
- Criteria of Good Test Norms

MEANING OF NORM-REFERENCING AND CRITERION-REFERENCING

An individual's performance in any psychological and educational test is recorded in terms of the raw scores. Raw scores are expressed in terms of different units, such as the number of trials taken within a specified period to reach a criterion; the number of correct responses given by the examinees; the number of wrong responses given; the total time taken in assembling the objects; and the like. All these raw scores convey no meaning in themselves. For example, when A has a score of 40 on the arithmetical reasoning test and a score of 30 on the history test, would it mean that A is superior or inferior or equivalent on the arithmetical reasoning test to the history test? In the absence of some interpretative data, it is difficult to answer. Usually, there are two reference points that are applied in interpreting the test scores (that is, in interpreting what the scores tell us about the examinees with respect to the characteristics being measured).

The first way is to compare an examinee's test score with the score of a specific group of examinees on that test. This process is known as *norm-referencing* and a test based on norm-referencing is called as norm-referenced test. In simple words, a norm-referenced test compares each person with a norm. The purpose of a norm-referenced test is to classify the persons from low to high across a continuum of ability or achievement. Such test, therefore, uses a representative sample of individuals—the norm group—as its interpretative framework. The researchers might want to classify individuals in this way for the purpose of selection to a specialized course or placement in any remedial or gifted programme. When raw scores are compared to the norms, a scientific meaning emerges. Norms may be defined as the average performance on a particular test made by a standardization sample. By a standardization sample is meant a sample which is the true representative of the population and takes the test for the express purpose of providing data for comparison and subsequent interpretation of the test scores. For adequate representation, the sample must include a cross-sectional representation of different parts of the population. In order to compare the raw scores with the performance of the standardization sample, they are converted into what is called 'derived score'. There are two reasons for such conversions. First, the derived scores provide occasion for direct comparison of the person's own performances on different tests, because they are expressed in the same units for different tests (whereas raw scores cannot be expressed in the same unit for different tests).

Second, the derived scores denote the person's relative position in the standardization sample, and therefore, his performance may be evaluated in comparison to other persons (raw scores do not provide this facility).

The second way of interpreting a test score is to establish an external standard or criterion and compare the examinee's test, scores with it. This process is known as *criterion-referencing*. In a criterion-referenced test, there is a fixed performance criterion. If an examinee passes some predetermined number of items (the criterion) or answers them correctly, it is said that he is capable of the total performance demanded by the test. Thus the criterion-referenced test may be defined as one in which the test performance is linked or related to some behavioural measures or referents (Glaser 1963). One question which arises here is: From where do the criteria come with which a test is referenced? According to Cox & Vargas (1966), a major criterion for referencing a test is the training, which results in an increase in the skill or proficiency. Suppose a group of examinees is given training, and the group is measured by a test before training and after training. If after training the scores are better than the pretraining scores, the test can be considered as sensitive to the results or outcomes of the training. In other words, the test scores or results can be interpreted as an indication of increased skill or proficiency. Such a test constitutes an example of the criterion-referenced test. The important features of a criterion-referenced test are as follows:

1. The test is usually based upon a set of behavioral referents, which it intends to measure.
2. The test represents the samples of actual behaviour or performance.
3. The performance on a criterion-referenced test can be explained in terms of predetermined cut-off scores.

It should be noted that a test cannot be automatically called a criterion-referenced test simply because it has no norms. Thus a criterion-referenced test describes the specific types of skills, tasks or knowledge that the examinee or test-taker can demonstrate. Such tests are used to compare the examinee's accomplishment to a predefined *performance standard*. Here the examinee's result is compared with the predetermined standard, and no comparison is made with other examinees. In fact, it is entirely possible for all examinees to exceed the predetermined standard. Thus, the central focus in criterion-referenced tests is on what the examinee *can do* rather than on *comparisons* to the performance levels of others. In this way, the criterion-referenced tests identify an examinee's relative mastery (or nonmastery) of specific competencies. Today, these kinds of tests are very popular in education systems.

A criterion-referenced test differs from a norm-referenced test. A major difference between the two is the manner in which the test content is chosen. As we know, in a norm-referenced test, the items are selected in such a manner that they provide maximal discrimination among respondents along the dimension being measured. Here well-defined psychometric principles are used to identify ideal items according to the difficulty level, correlation with total scores and several other properties.

The specific differences between a norm-referenced test and a criterion-referenced test are as under:

(i) The purpose of the criterion-referenced test is to compare the examinees' performance to some predetermined criterion or standard. On the other hand, the purpose of the norm-referenced test is to compare the examinees' performance to one another.

(ii) Items of the criterion-referenced test reflect narrow domain of skills with real-world relevance whereas items of the norm-referenced test reflect broad domain of skills with indirect relevance.

(iii) In criterion-referenced tests, scores are usually expressed as a percentage with a pre-determined passing level whereas in norm-referenced test scores are usually expressed as standard score, percentile or grade equivalent.

(iv) Another feature that makes the criterion-referenced test distinct from norm-referenced test is that in criterion-referenced test, most items are of similar difficulty level whereas in norm-referenced test items vary widely in difficulty level.

STEPS IN DEVELOPING NORMS

Developing norms is certainly a very difficult task. However, this difficulty can be minimized if we follow the proper steps in developing norms. The following are the three important steps in developing norms.

1. Defining the target population
2. Selecting the sample from the target population
3. Standardizing the conditions

These steps may be discussed as follows:

1. Defining the target population:

The first step in developing norms is to define the composition of the target group. The test is intended to be used for a particular type of person or group of persons. The composition of the target group (also called normative group) is determined by the intended use of the test. Let us suppose that the test constructor has constructed *Test of English as Foreign Language* (TOEFL). Obviously, this test is intended for those students whose native tongue is not English but who plan for studying abroad where the medium of instruction is English. Thus, for TOEFL the target population will consist of these students who are relevant and appropriate. For TOEFL, population of PhD candidates will be an example of inappropriate one.

2. Selecting the sample from the target population:

When the test constructor has defined the target population, he proceeds to select a representative sample from each of the target population or group. To make the selected sample representative of the target population, a cross-sectional representation of the target population must be made. A cross-sectional representation is one in which people from all sections of the target population are represented. For ensuring representative character of the sample, various techniques of sampling are employed. Generally, for constructing norms a larger sample is preferred. For constituting a larger sample, a completely random sampling technique is the best one; but due to its impracticability, this technique is seldom employed. Generally, cluster sampling or its variation is preferred.

3. Standardizing conditions for proper implementation of test:

Unless conditions of test administrations are standardized, valid and proper comparisons of individual test scores to test norms are impossible. Therefore, factors like adequate sound control, lighting, ventilation, temperature of working space and the like must be properly controlled. Above all, factors like test timing, test security, adherence to test-manual direction and assuring that the examinees work on proper test sections are most important for standardizing the conditions of working space. Without these standardization procedures, norms cannot serve as a useful comparative device.

These are the important steps in developing norms of a test. For norms to be a useful comparative device, these steps must be covered thoroughly.

TYPES OF NORMS AND TEST SCALES

Ordinarily, the derived scores may be divided into four common types and depending upon each of these four scores, there are four types of norms. The four commonly derived scores are age scores, grade scores, percentile scores and standard scores (Lyman 1963). Accordingly, there are four types of norms—age norms, grade norms, percentile norms and standard-score norms. A detailed critical discussion of each of these is given below.

Age-equivalent Norms

Age-equivalent norms are defined as the average performance of a representative sample of a certain age level on the measure of a certain trait or ability. If, for example, we measure the weight of a representative sample of 10-year-old girls of the state of Bihar, and find out the average of the obtained weight, we can determine the age norms for the weights of 10-year-old girls.

Age norms are most suited to those traits or abilities which increase systematically with age. Since most of the physical traits like weight, height, etc., and cognitive abilities like general intelligence show such systematic change during childhood and adolescence, age norms can be more appropriately used for these traits or abilities at the elementary level.

There are some disadvantages of age norms.

1. Age norms lack a standard and uniform unit throughout the period of growth of physical and psychological traits. As pointed out earlier, age norms are suited to traits or abilities which show a progressive growth with advancement of age. Consider the example of general intelligence. The growth in the level of general intelligence from age 8 to 9 is in no way equal to the growth from 3 to 4 and 14 to 15. This is true because the growth at the earlier level is faster than the growth at a later level and it almost comes to a halt after 16 to 17. Not only this, this problem becomes further aggravated due to the fact that even at a particular age level the rate of growth of general intelligence (or traits like height and weight) is not uniform for all children and sometimes it varies to a great extent. In such a situation, we are forced to view with skepticism the meaning of equality of age units used in age norms.

2. Another problem in age norms arises from the fact that the growth rate of some traits are not comparable. For example, progress in maze learning does not ordinarily take place after adolescence but progress in vocabulary continues even after adolescence. In such a situation, the age norms for these two traits cannot be compared.

3. A trait like acuity of vision cannot be expressed in terms of age norms because this trait does not exhibit progressive change over the years. Many other personality traits would also fall into this category.

Grade-equivalent Norms

Like age-equivalent norms, grade-equivalent norms are defined as the average performance of a representative sample of a certain grade or class. The test, whose norms are being prepared, is given to the representative sample selected from each of the several grades or classes. After that, the average performance of each grade on the given test is determined and then grade equivalents for the in-between scores are determined arithmetically by interpolation. This average performance is known as the grade-equivalent norms. Thus, if the average number of items done correctly on the arithmetic test is 30 by the sixth grade that constitutes a representative sample then a raw score of 30 becomes a grade norm of six. In this way, the grade-equivalent norms indicate the grade levels at which the performance of representative groups is average.

Grade-equivalent norms also have some limitations.

1. Grade-equivalent norms of the same student in different subjects are not comparable. For example, the grade-equivalent in social studies of a student cannot be compared with the grade-equivalent in arithmetic of the same student because everyday life experiences outside the school may contribute to the knowledge of social studies whereas knowledge of arithmetical concepts is primarily dependent upon formal training in the class.

2. Grade-equivalent norms assume that all students of a class or grade have more or less similar curriculum experiences. This assumption may be true in the elementary classes but it may not be true for higher classes.

3. The grade-equivalent norm is not suited to those subjects in which there occurs rapid growth in the elementary class and a very slow growth in the higher classes. For example, in spelling and arithmetic there occurs rapid growth in the elementary grades but growth is slow in the higher classes.

Despite these limitations grade-equivalent norms are common, particularly among achievement tests and educational tests. Such norms are also suited to intelligence tests.

Percentile Norms (or Percentile-rank Norm)

Before the meaning of percentile norms is explained, it is essential to look at the related terms: *centiles, deciles* and *quartiles*. As we know, a scale of percentile rank consists of 100 units and the percentile rank of a score represents the percentage of cases in a distribution that had scores *at or lower than the one cited*. When a score is identified by its percentile rank, the scale is called as a percentile. Suppose, for example, that Mohan has a score (X) of 53 on an achievement test and that 70% of the class (to which Mohan belongs) had exactly a score of 53 or lower. Thus, Mohan's score $X = 53$ has a percentile rank of 70 or PR_{70} and his score would be called as 70th percentile. Thus, percentile rank refers to a *percentage* and that percentile refers to a score. A *centile* is equivalent to a percentile and is the percentage rank that divides the distribution into 100 equal parts. Thus 30th centile is specified as C_{30} and is equivalent to 30th percentile. A *decile* is the percentage rank that divides the distribution into 10 equal parts. The 7th decile (specified as D_7) is equivalent to the 70th centile (percentile). The 1st decile is the 10th centile, etc. A *quartile* is a percentile rank which divides a distribution into 4 equal parts. There are 3 quartiles in any scale of percentile ranks and they include 1st quartile (Q_1) which is equivalent to 25th centile, the 2nd quartile (Q_2) which is equivalent to the 50th centile or 5th centile and the 3rd quartile (Q_3) which is equivalent to the 75th centile.

Percentile norms are the most popular and common type of norms used in psychological and educational tests. Such norms can be prepared for either adults or children and for any type of tests. A percentile norm indicates, for each raw score, percentage of standardization sample that falls at or below that raw score. To illustrate, suppose Mohan has a score of 26 on the mechanical reasoning test and if 40% of the standardization sample secures below the score of 26 then Mohan has percentile rank of 40 or PR_{40} and percentile scores of 26. Percentile norms, thus, provide a basis for interpreting an individual's score on a test in terms of his own standing in a particular standardization sample. If the percentile norm is to be meaningful, it should be based upon a sample which has been made homogeneous with respect to age, grade, sex, occupation and other factors, otherwise separate tables for percentile norms for each age, grade, sex and occupation should be prepared.

Sometimes, percentile norms are also reported through the ogive—a graph which shows a cumulative percentage of scores falling below the upper limit of each class interval. The graph enables a person to read his score (plotted on the X axis) on the basis of percentile rank or cumulative frequency percentage (plotted on the Y axis); or the percentile rank can be read on the basis of the person's score. Some of the important advantages of percentile norms are that they are easy to construct, easy to understand, and even an untrained person can freely use them. Despite these advantages, such norms have two distinct limitations.

1. Laymen as well as skilled persons sometimes fail to distinguish between the percentile and the percentage score, and the obvious result is confusion. But the two should not be confused because percentile is a derived or converted score usually expressed in terms of percentages of persons whereas the percentage score is a raw score expressed in terms of percentages of items carefully answered or solved.

2. The biggest limitation of the percentile norms is that of inequality of units throughout the percentile scale. The common observation is that most of the raw scores obtained in psychological and educational measurement approximate the normal distribution and when they are expressed in terms of percentile norms, the single raw score difference at or near the middle (such as percentile rank or PR in the range of 40–60) may lead to a change of several percentile rank points, whereas the large raw score difference at the extremes of the percentile scale (such as at PR of 90 or PR of 10) may produce a very small percentile rank difference. In other words, raw score differences at or near the middle of the percentile scale produce exaggeration in percentile ranks whereas the raw score differences at either extremes of the percentile scale produce a very small change in percentile ranks. Thus percentile scores are the nonlinear transformation of raw scores. The above point has been illustrated in Figure 7.1, which reveals among other things that at the centre or mean of the distribution, 5% of the cases lie between PR of 50 and 55. This 5% of the cases have been shown by a tall narrow bar. The same 5% cases at the extreme lying between PR of 90 and 95 make a broad low bar. Thus, the same number of percentile-rank points covers a wider range of scores at the extreme and a narrower range of scores at the centre or median of the distribution. In other words, a smaller (or narrower) raw-score difference (shown by tall narrow bar) at the median is definitely producing an exaggeration in percentile ranks (that is, difference of 5 percentile ranks) whereas only the bigger raw-score differences (shown by the broad low bar) at the extreme are able to produce changes in the percentile ranks to that effect. This means that the raw score differences at the extremes are highly shrunk in percentile transformation. Thus, percentile units of the percentile norms are systematically unequal.

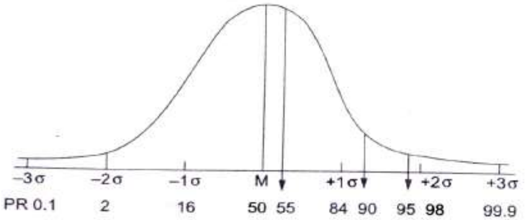

Fig. 7.1 Selected percentile ranks in normal distribution

3. Percentile norms indicate only the person's relative position in the standardization sample. They convey nothing regarding the amount of the actual difference between the scores.

Standard Score Norms

A norm which is based upon a standard score is known as a standard score norm. The reason why, one needs standard score norms in place of percentile norms is that here units of the scale are equal so that they convey the same meaning throughout the whole range of the scale. In this way, standard score norms remove one of the serious problems of inequality of units common among the percentile norms.

In order to understand the standard score norms, it is essential to first know the meaning of the standard score itself. Standard score, like the percentile score, is a derived score. It has a specified or fixed mean and fixed standard deviation. There are several types of standard scores, such as *z* score (also known as sigma score), T score, stanine score, deviation IQ, etc. Of these, the *z* score is very popular and has been frequently used in preparing norms. In fact, some writers use the term "standard score" interchangeably with the *z* score but in reality, this should not be so because the *z* score is just one of several types of standard scores.

Why are standard scores needed? Standard scores are needed primarily for two reasons. Firstly, when the performance of the same person on different tests is to be compared, it is best done through converting the raw scores into standard scores. Secondly, standard scores have equal units of measurement and their size does not vary from distribution to distribution. Hence, they are frequently used in interpreting test scores.

Raw scores can be converted into standard scores through two methods—linear transformation and normalized transformation. When raw scores are linearly transformed into standard scores, all characteristics of the original distribution of the raw scores are retained without any change in the distribution. As such, all statistical computations which can be carried out with original distribution of raw scores can be duplicated with the standard scores. The *z* score is an example of a linearly transformed standard score. A *z* score indicates how many standard deviations of a score lie below or above the mean of distribution. A *z* score is a standard score which has a mean of zero and standard deviation of 1. It is computed by subtracting the mean from the individual's raw score and dividing the same by the standard deviation (SD) of the distribution. In terms of an equation, a *z* score may be read as under:

$$z = \frac{X - M}{\sigma} \tag{7.1}$$

where z = z score or sigma score; X = raw score; M = mean of the distribution; σ = standard deviation of the distribution.

To illustrate, suppose Mohan has obtained a score of 60 on a test and the mean and SD are 30 and 15 respectively. Then,

$$z = \frac{60 - 30}{15} = \frac{30}{15} = 2$$

This means that Mohan's score is 2 sigma or standard deviations above the mean of the distribution. Likewise, if another person has obtained a score of 20 on the same test then

$$z = \frac{20 - 30}{15} = \frac{-10}{15} = -0.66$$

This means that his score is 0.66 standard deviation below the mean of the distribution. Likewise, if another person obtains a score of 30, it means his score is exactly equal to a *z* score of zero.

z score is often used in making comparisons. In fact, the transformation of raw scores into z scores is very useful when the researcher wants to compare scores from two different distributions. For example, suppose Bobby earned a score of 60 on the test of psychology. Further, suppose that mean was 50 and standard deviation (σ) was 10. In the geography test, Bobby's score was 56 and mean was 48 and σ = 4. Now the question is: In which class does Bobby have a better standing? At a glance, Bobby's psychology score is higher than his score in geography. Also, a look at deviation score shows that Bobby is 10 points above the mean in psychology and only 8 points above the mean in geography. Can we conclude that she performed better in psychology than in geography? Reality is that we can't compare her psychology score to geography score because these scores come from *different distributions*. For this, a score of 60 in psychology means something different from a score of 56 in geography. Any direct comparison between these two test scores would be like the proverbial comparison of *apples to oranges*. In fact, for determining the class in which Bobby did better, it is essential that the distribution of both classes must be *standardized* to make them similar so that meaningful comparisons can be made. Therefore, both scores are converted into z score as follows :

$$z = \frac{X - M}{\sigma} = \frac{60 - 50}{10} = \frac{10}{10} = +1.0 \text{ (for psychology)}$$

$$z = \frac{X - M}{\sigma} = \frac{56 - 48}{4} = \frac{8}{4} = +2 \text{ (for geography)}$$

Now it is clear that Bobby's z score for geography was +2, which means that her test score is two standard deviation above the class mean. On the other hand, her z score for psychology is +1.00 or one standard deviation above the mean. In terms of relative class standing, Bobby is doing much better in geography than in psychology because z scores describe relative positive within a distribution.

There are two merits of z score. First, z score represents the most precise way of indicating position in distribution. Second, although it does not sound good when computing average of ranks or percentile ranks, it is perfectly legitimate when z scores are being used. For example, suppose a person has a z score of 1.25 on the mathematics test and a z score of –0.5 on a verbal comprehension test, the number of standard deviations away from both means, on the average, would be [1.25 + (–.5)]/2 = 0.75/2 = 0.375 standard deviation above the mean of X's.

There are some difficulties associated with the use of z scores. First, a z score may be obtained in plus and minus form as it is obvious from the above example. z scores in minus are cumbersome to handle. Second, z scores may be expressed in decimal points which tend to further complicate handling them. In order to circumvent the first difficulty the score is made larger by adding to each z score a constant, such as 50 or 100, so that the minus signs are omitted. The difficulty caused by decimal points is solved by using a larger standard deviation, that is, by multiplying every z score by 10 or 20. Such linear transformation of a z score into a new standard score is very common, particularly the system of having a mean of 50 and SD of 10. The formula for such transformation of the z score is shown below.

$$\text{Standard score} = z\,(\text{new standard deviation}) + \text{new mean} \qquad (7.2)$$

To illustrate the use of this equation, let us consider the examples given in Table 7.1 in which the raw score has been converted into a z score and the z score has been converted into a standard score with a mean of 50 and SD of 10.

Table 7.1 Transformation of z score into new standard score

Students	Raw Score	z Score	Standard Score
			Mean = 50; SD = 10
A	72	1.2	62
B	40	–2	30
	Mean = 60		
	SD = 10		

In Table 7.1, A's raw score of 72 has been converted into a z score through Equation 7.1 and the z score has been further converted into a standard score with a mean of 50 and SD of 10 through Equation 7.2. Thus,

$$z = \frac{72 - 60}{10} = \frac{12}{10} = 1.2$$

$$\text{Standard score} = 1.2\,(10) + 50 = 12 + 50 = 62$$

$$z = \frac{40 - 60}{10} = \frac{-20}{10} = -2$$

$$\text{Standard score} = -2\,(10) + 50 = -20 + 50 = 30$$

One of the important uses to which the above linearly transformed standard score can be put is in the comparison of the same person's performance on different tests. Table 7.2 serves as an illustration of this. In this table, the raw scores (shown in Part I) have been converted into standard scores with a mean of 50 and SD of 10 (shown in Part II) through Equation 7.2.

Table 7.2 Linear transformation of raw scores into standard scores*

Part I: Raw Scores

Students	Psychology	Sociology	Education
A	120	90	55
B	100	70	55
C	110	80	45
D	130	60	30
Mean	100	60	35
SD	20	10	5

* Raw scores of only four students have been shown.

Part II: Standard Score
Mean = 50; SD = 10

Students	Psychology	Sociology	Education	Average
A	60	80	90	76
B	50	60	90	66
C	55	70	70	65
D	65	50	60	58

Part I of Table 7.2 shows raw scores of only four students out of eight students, and Part II shows linearly transformed standard scores. At the bottom of Part I are given the mean and SD of the distribution. If you look at Part I, it conveys little meaning. What is the overall performance of student A on the three tests? On which test did he perform the best or worst? Which of the four students had the best overall performance? None of these questions can be answered. But when these scores are transformed into a standard score, having mean = 50 and SD = 10, these questions can be easily answered. Student A had a standard score of 60 in psychology, which is I standard deviation (mean = 50 and SD = 10) above the mean; in sociology he is 3 standard deviations above the mean and in education he is 4 standard deviations above the mean. Accordingly, it can be said that A's performance is the best in education; excellent in sociology and slightly above average in psychology. His overall performance over these three tests is 76 which is 2.6 standard deviations above the mean (that is 50). In a similar manner, the performance of each student can be compared. Of the four students the best overall performance is that of A because his performance is 2.6 standard deviations above the mean while the performance of B, C and D is 1.6, 1.5 and 0.8 standard deviations respectively above the mean.

Normalized Standard Score

So far we have discussed the linear transformation of raw scores to standard scores through the z score. Such standard scores have all the properties of the original distribution of raw scores, that is, standard scores are related to raw scores through a linear function. One general demerit of linearly transformed standard scores is that they can be compared only when they have been derived from distributions that have approximately the same form. If one distribution is skewed and another is normally distributed then the two standard scores cannot be compared. In such a situation where the comparison of the standard scores derived from different types of distribution becomes necessary, a direct solution is to convert raw scores (which have a skewed distribution) into standard scores, which are normalized and then, compare the two standard scores. Such standard scores are called normalized standard scores. Marshall & Hales (1972, 120) have defined normalized standard scores as scores which have been "adjusted to produce a normal frequency distribution and converted to a standard base with a preassigned mean and standard deviation." Thus normalized standard scores are the standard scores, which are expressed in terms of the *normal distribution*. Normalized standard scores can be expressed in the same form as linearly derived standard scores, that is, with a mean of zero and standard deviation of 1. A normalized standard score can therefore, be interpreted in a similar way. A normalized standard score of +1 indicates that the examinee has surpassed 84% of his group and –1 indicates that he has surpassed only 16% of his group. There are three common normalized standard scores—T scores, stanine scores and the deviation IQ.

Like linearly derived standard scores, normalized standard scores can be expressed in several forms. The T score was first devised by McCall (1922) and named after E L Thorndike, an important leader in educational measurement. T score is defined as a standard score which is based upon a mean of 50 and standard deviation of 10 as well as upon the *shape of the normal distribution curve*. The T-score scale is based upon the normalized *z*-score scale, which generally extends from –3 to +3 standard deviations with decimal fractions at various points in between. The T-score scale has a range of 20 to 80 in most distributions (Figure 7.2), when raw scores are transformed into T scores and their distribution automatically takes a shape which approximates a normal curve. In terms of a formula, T score is equal to:

$$T = z10 + 50 \tag{7.3}$$

where *z* refers to the *z* score and is computed by Equation 7.1.

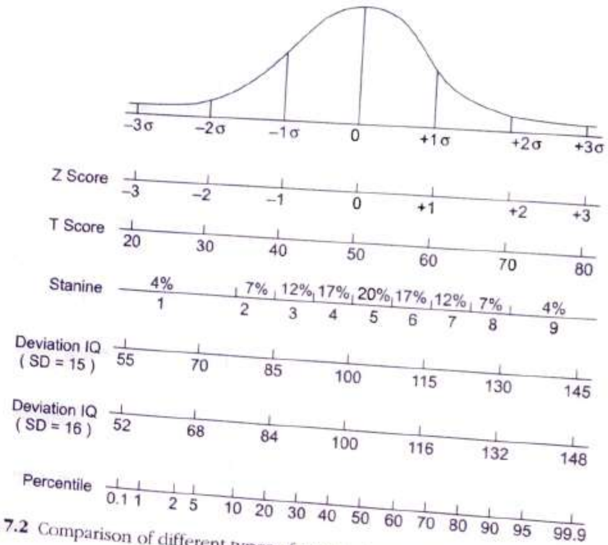

Fig. 7.2 Comparison of different types of standard scores in a normal distribution

Equation 7.3 converts a raw score (not grouped into frequency distribution) into a T score. But when the raw scores have been grouped into a frequency distribution, a direct table[*] for their conversion into T scores is available.

Another normalized standard score is the stanine score which was developed by the United States Air Force during World War II. The term 'stanine' is a contraction of 'Standard nine' and it has scores expressed in digits ranging from 1 to 9. The mean of these scores is 5 and the standard deviation is 1.96 or approximately 2 (Figure 7.2). When raw scores are transformed into stanine scores, they automatically take a shape approximating the normal curve. As a matter of fact, stanine scores are the condensed scores on the C scale. In the C scale, there are 11 score points ranging from 0 to 10 with the mean lying exactly at 5. For computational facilities with computer punched-card records the two points at both the extremes (that is, 0 on the lower end and 10 on the higher end) are combined, thus leaving only a nine-point scale (called the stanine scale). A variant of the stanine scale is the sten scale proposed by Canfield (1951) where there are 10 units—5 units above and 5 units below the mean.

Raw scores can be transformed into the stanine scale by arranging them in order of size and then giving the percentage of each stanine score points according to the normal distribution curve (Figure 7.2). The first stanine covers 4%, second stanine covers 7%, third stanine covers 12%, fourth stanine covers 17%, fifth stanine covers 20%, sixth 17%, seventh 12%, eighth 7%, and ninth 4% of the total cases. When, for example, there are 300 scores earned by 300 students on a test, then the lowest 12 scores (4% of 300) would receive a stanine score of 1; the next 21 scores would receive a stanine score of 2, and so on.

Another normalized standard score is the deviation IQ score. Such a normalized standard score has a mean of 100 and a standard deviation of about 16 (Figure 7.2). The size of standard deviation is, however, dependent upon the intelligence test being used. Usually, standard deviation varies from 12 to 18 but a uniform standard deviation of 15 or 16 is much in vogue. With variations in standard deviation units there will also be variations in percentages of cases in normal distribution and accordingly, variations in score points (Figure 7.2). With a mean of 100 and SD of 16, the deviation IQ of 132 would be 2 standard deviations above the mean and the deviation IQ of 84 would be –1 standard deviation below the mean. But when an SD of 12 is taken, the deviation IQ of 132 would be 2.66 standard deviation above the mean and the deviation IQ of 84 would be –1.33 standard deviation below the mean.

The deviation IQ score and the traditional IQ score (called the ratio IQ score found by dividing the mental age by the chronological age) has not much similarity. The students should not be confused regarding the meaning of these two terms because the deviation IQ is not really IQ, rather it is a normalized standard score, nor is it calculated in a manner in which the ratio IQ is calculated. As a matter of fact, many experts have objected to the term 'deviation IQ' because it is a misleading one. The only justification in calling such standard scores as the deviation IQ is that they can be *interpreted* like IQ's provided the chosen SD on the test is equal to the previously known IQs. The first test to express its scores in terms of the deviation IQ was the Wechler Intelligence Scale in which the mean is 100 and the SD is 15. Nowadays, most of the intelligence tests use the deviation IQ score rather than the ratio IQ score in order to circumvent the problems posed by the latter. For a discussion of these problems, students are referred to Chapter 9.

Local and Subgroup Norms

Sometimes local and subgroup norms are constructed to suit the specific purpose of the test. Local norms are defined as those norms, which are derived from representative of local examinees rather than from a national sample. A test having norms derived from Hindi-speaking people of Bihar only (rather than Hindi-speaking people of whole India) will be an example of local norms. Subgroup norms are those norms that consist of scores obtained from an identified subgroup which can be formed with respect to sex, ethnic background, geographical region, urban vs. rural environment, socio-economic level and many other similar factors.

---

[*] The interested reader may consult Guilford & Fruchter (1973, 469).

Which of these two norms is more beneficial? The answer to this question depends upon the purpose of testing. For example, local norms for standardized intelligence test may be considered superior in predicting the children's competence within their nonschool environment. However, such norms may not be much effective in predicting the children's competence in mainstream school instructional programme of the whole country.

COMPUTER APPLICATIONS IN PSYCHOLOGICAL TESTING AND ASSESSMENT

The application and use of computers in psychological testing and psychological assessment has been a recent development. For testing and assessment, computers can be used in two basic ways: (a) In creating new tasks and pickups measuring abilities that traditional procedures can't tap (b) In administering, scoring and interpreting traditional tests.

An early inspiration for using computers to generate tasks that one can't present through traditional methods, in fact, came from the application of psychophysical and signal-detection procedures. In these procedures, a stimulus or signal is presented and the subject is required to report his perception. These psychophysical and signal-detection procedures are being applied to ability testing (Jensen1982; Nettlebeck 1982). Backward masking tasks (in which the target stimulus is followed by noise stimulus known as *mask*) and reaction time are commonly used to measure the speed, capacity or efficiency of information processing (Saccuzzo, Larson & Rimland 1986). The prominent assumption is that variations among individuals who differ in psychometric intelligence tend to reflect different information-processing capabilities (Hunt 1980; Jensen 1986). Test developers have adopted the reaction time and backward masking tasks for presentation by personal computers (Saccuzzo & Larson 1987; Saccuzzo, Larson & Rimland 1986). Other investigators, with the help of standard software, can easily verify results from psychophysical and signal-detection procedures.

Still another application of computers is in the computerization of tests. Currently, there are numerous programmes for administering, scoring and even interpreting a host of tests (Baker, 1989; Gutkin & Wise 1991). At the simplest level, those tests which are meant for group administration are now adapted for *computer scoring*. Apart from this, there is also increasing availability of *computer disks* which can be employed by test users to score tests on their own computers. The ASSIST programme developed by American Guidance Science is one such example. At the more complex stage, *narrative computer interpretation* of the test scores are available for certain types of tests. In such an approach, verbal statements with particular pattern of test responses are prepared by computer programme associates. This approach has been widely used in personality and aptitude tests. In MMPI, the test users may obtain computer printouts of the diagnostic and interpretative statement about their personality together with the numerical scores.

Individualized interpretation of test scores is also done by what is called *interactive computer systems* in which the individual is in direct contact with the computer by means of a response pattern and thus, engaged in a dialogue with the computer (Katz 1974). The reality is that in this technique, test scores are usually incorporated in the computer together with other information provided by the person. Subsequently, the computer combines all the available information about the individual with stored data particularly about his educational and occupational background. It utilizes all relevant facts and relations in answering the individual's questions and also helps him in reaching certain decisions. A very good example of such an interactive computer system is SIGI (System for Interactive Guidance Information).

Several researchers have reported that the presentation of the test content and scoring of the test by computer does not appreciably reduce reliability or validity of test scores (Lee, Moreno & Sympson 1986). However, certain applications of computers may lead to misuses and misinterpretation of test scores (Butcher 1985; Matarozzo 1986). For guarding against such hazards, considerable attention has been given to developing important guidelines for computer-based testing. AERA, APA and MCME together have developed several standard guidelines for computer-based testing called *The Testing Standards*.

Today most of the experts agree with the view that pertaining to computerized testing, score *comparability* and *narrative interpretive scoring* are of major concern which deserve special attention. Mazzeo et al. (1991) have pointed that when the same test is administered in computerized mode and in traditional printed mode, the comparability of the scores need to be carefully investigated. The reliability and the validity of the test may also vary. Thus it can be concluded that unless the two modes are shown to yield fully equated test forms, the same set of norms may not be appropriate as well as applicable for both. Narrative interpretive scoring has also aroused some concern because sometimes it has been observed that adequate information is not provided to test users for evaluating the reliability, validity and other technical properties of the interpretive system. Likewise, at other times it is also not clear how the interpretative statements were derived from the scores. In the absence of such guidelines, narrative interpretive score reports no longer remain a meaningful affair.

The potential contribution of computers to psychological assessment is impressive (Gutkin & Wise 1991; Moreland 1992). Farrell (1992) has recently identified several major applications of computers to psychological assessment, specially to cognitive-behavioural assessment. These applications are (i) collecting self-report data, (ii) directly recording behaviours, (iii) coding observational data, (iv) training, (v) organizing and synthesizing behavioural assessment data, (vi) analyzing behavioural assessment data, and (vii) supporting decision making.

Computer technologies have also led to the development and rapid growth of many new instruments for the assessment of cognitive functions which are extensively being used in clinical neuropsychology as well as in the study of attention disorders and learning disabilities (Stoloff & Couch 1992). Computers allow us not only to vary task presentation conditions more precisely so as to assess performance on different task components but also allow recording and evaluation of response parameters such as timing in the way not possible by the traditional test. The Microcog: Assessment of cognitive functioning is one of the best examples of recently developed computer-administered battery which is designed to screen for likely signs of cognitive impairment in adults (Powell & Whitla 1994).

However, some factors continue to impede the widespread application of computers in psychological assessment. *First*, the lack of acceptance by some practitioners is one of the major factors. *Second*, evaluation of software is another obstacle (Farrell 1992). Most of the vendors are reluctant to make their products available for review. Consequently, potential users don't have proper and sufficient information to evaluate the quality of the software. *Third*, most systems of computer-assisted test interpretation tend to combine clinical and statistical procedures. This specific mix of quantitative data and clinical judgment varies among the systems in the same way as the technical quality of the database and the clinical expertise of the judgment. Apart from this, the needed information to evaluate a particular system is often unavailable because of proprietary concern.

Despite these obstacles, it is hoped that greater use of computers in both psychologica testing and psychological assessment will be witnessed in future.

CRITERIA OF GOOD TEST NORMS

We have seen that test norms are of different types. Regardless of the type of test norms, it is essential that their adequacy should be established. Test developers have set some criteria of good qualities that norms should possess. The following are main criteria for judging whether or not the test norms possess good qualities:

1. *Test norms should be representative:* One of the important characteristics of test norms is that they should be a true representative of the group with which comparison is desired. Ideally, test norms should be based on a random sample taken from the population they represent. This is extremely difficult as well as expensive. Therefore, the matter has to be settled. At the minimum, all the subgroups of the population such as boys and girls, rural and urban areas, socio-economic status, caste groups, schools of varying size and so on, must be adequately represented in the sample.

2. *Test norms should be relevant:* Another good quality of norms is that it should be relevant. Since norms are based on various types of groups, they must be meaningful in light of the concerned group. Some test norms are based upon national sample whereas some are limited to samples taken from a limited geographical region or state. The variety of groups available for comparison necessitates the study of the norm sample before using any table of norms. If the researcher wants to compare sample or student with a general reference group for diagnosing the strengths and weaknesses in different areas, national norms may be a good decision. But when the researcher is trying to decide such things as which students should be placed in a highly select group, which ones should be encouraged to pursue engineering course, which ones should be encouraged to pursue medical course, national norms are less fruitful and for such decision, norms for each specific groups are needed. A student might have an above-average aptitude and achievement when he is compared with students in general, but he may fall short of the ability needed to succeed in any accelerated group.

3. *Test norms should be up to date:* Good test norms should be up to date and they should not be based upon past uses. When achievement and education are going up, a given raw score will produce a higher percentile rank when compared to old norms than when compared to new norms. Therefore, old norms should be discarded and new norms should be preferred.

4. *Test norms should be comparable:* Another good quality of test norms is that they should be comparable. When the researcher wants to directly compare scores from different tests such as when he compares aptitude and achievement test scores to identify underachievers or when he makes profile comparisons of test scores to identify the strength and weaknesses of a student, comparability of test norms is needed because only comparable test norms can help the researcher in doing this desired function. A comparability of norms is assured when the norms of all tests have been developed on the same population. Whenever the researcher wants to compare scores from different tests directly, he should check the manual to ascertain whether the norms are based on the same group or whether they have been made comparable by any other means.

5. *Test norms should be properly and adequately described:* Test norms should provide information about the norm group, the norming procedures and other relevant factors. Such information is needed and only then, one is able to judge the appropriateness of test norms for any particular purpose. Test norms should provide at least the following information:

(i) What is the method of sampling?
(ii) What are the number and distribution of cases in the norm sample?
(iii) What are the characteristics of the norm group especially regarding age, sex, socio-economic status, geographical location, education level, caste, social group, etc.?
(iv) Whether standard conditions of administration and motivation were maintained during testing.

☑ Review Questions

1. What do you understand by norms of a test? Why are norms needed for a psychological test?
2. Make a distinction between age norms and grade norms.
3. What is meant by percentile norms? Discuss the advantages and disadvantages of percentile norms.
4. What is the standard about standard score? Discuss the research utilities of different types of standard-score norms.
5. Discuss the major computer applications in psychological testing and assessment.
6. Discuss the major criteria for a good norm.

◆

8

RESPONSE SET IN TEST SCORES

CHAPTER PREVIEW

- Meaning of Response Sets
- Types of Response Sets
    - Deviation Set
    - Acquiescence Set
    - Faking-Good and Faking-Bad
    - Semantics
    - Evasiveness
    - Tendency to work Speedily
    - Tendency to Guess When Uncertain
    - Cautiousness
    - Extremeness
- Implications of Response Sets
- Methods to Eliminate Response Sets
    - Design of Test Items
    - Adapting Test Difficulty
    - Use of Correction Keys
    - Modifying the Directions
    - Use of Good and Appropriate Scoring Formulas
    - Use of Easy Multiple-choice Items

MEANING OF RESPONSE SETS

There are several factors which tend to invalidate test scores. Response bias is one of them. When a response to any test item tends to be altered by the examinees in a way that it indicates something other than that which the test is intended to measure, it is commonly called the response bias (Guilford 1954). A response set is a kind of response bias which occurs as a result of assuming certain mental sets by the examinees. For example, the examinee may assume for reasons best known to him that only the 'Yes' alternative of 'Yes-No' is correct in a particular test, and accordingly he may answer all items in terms of 'Yes'. Cronbach (1946, 476) has defined response set as "the tendency causing a person to give different responses to items than he would when the same content is presented in a different form." To illustrate the definition of Cronbach, suppose that three examinees—A, B, and C—have reported identically on a two-alternative item of a personality inventory. But they might respond differently when the same content is provided by the four-alternative item of a personality inventory (a multiple-choice item). Assuming that these differences were not due to poor reliability, they may be said to be due to some kind of response set operating within all the three subjects. For example one subject might be guided by the social desirability tendency (faking-good), another by the social undesirability tendency (faking-bad) and still another by the impulsion set. Thus a response set tends to make the response

different from that which the test really intends to measure. Response sets are commonly identified in tests of ability, personality, interest, attitude scales, etc. There are several types of response sets. The following are the main response sets operating within the examinees.

TYPES OF RESPONSE SETS

Deviation Set

The deviation set was first proposed by Berg (1961). It refers to a tendency to give very unusual or uncommon responses to items in a test. Such a set is often found in the case of tests having nonverbal items such as the Rorschach test, the Thematic apperception test, etc.

Acquiescence Set

Acquiescence is another response set found often in case of the personality inventory and the interest inventory. It refers to a tendency to respond in some systematic way such as marking only 'Yes' or 'True' or 'Agree' (Yes saying) or only 'No', 'False' or 'Disagree' (No saying) on each item. The acquiescence is taken to be a continuous variable—at one end are examinees who consistently give 'Yes' or 'True' answers and at the other end are the examinees who consistently give 'No' or 'False' answers. Acquiescence sets have several important consequences on the total score of the test. If the test's 'Yes' or 'True' responses are keyed positively, guessing due to this set may enhance the total score more than the actual score. Likewise, if the test's 'No' or 'False' responses are keyed positively, guessing due to this set may enhance the total score more than the actual trait. In either case, a distorted picture of the examinees emerges.

Faking-Good and Faking-Bad

Faking-good, also known as the *social desirability tendency* is a very common response set in personality tests, service selection tests, etc. As its name implies, social desirability tendency refers to a tendency on the part of the examinees to give answers in a way that tends to create a favourable impression of themselves. Thus the faking-good response set helps the examinees in increasing their chances of being selected by painting a good picture of the self. Such a tendency is very frequent in those tests which assess socially desirable traits like emotional stability, punctuality, trustworthiness, responsibility, etc. A faking-good tendency has also been popular among tests on the basis of which the examinees are to be selected for a job or to be admitted to a school or a college. It has also been observed that under certain circumstances the examinees are more motivated to fake bad, that is, to select answers that may give an unfavourable impression or socially undesirable impression of the self. Such a tendency is known as the faking-bad or the *social undesirability tendency*. A patient who is to undergo psychotherapy may give such answer on an adjustment inventory which may show that he is more maladjusted. Likewise, a military officer who wants to return home, may give responses on a personality inventory which make him appear more mentally ill than he actually is.

What is indicated by the tendency to choose a socially desirable response in a personality inventory? The common-sense viewpoint suggests that the tendency of faking-good indicates a kind of deliberate deception on the part of the examinees. But empirical evidences are otherwise. According to Edwards (1957) who has done extensive researches on the social desirability variable, such tendency is nothing but a sort of facade effect or a tendency to "put oneself in a good front". He further revealed that the examinees are mostly unaware of such a tendency. Crowne & Marlow (1964) and Frederiksen (1965) held the view that the social desirability tendency was related to the need for social approval, self-protection, social conformity and the need for avoiding social criticism. Likewise, the social undesirability tendency (or faking-bad) may indicate the need for sympathy, attention and being helped by others in solving problems. Experimental evidences have indicated that when the examinees are *instructed* to fake good on a test, (that is, instructed to give responses in a manner that could make better show off), they are able to do that better and accordingly, improve their scores than when a neutral instruction is given to them. This has one interesting implication for the test. The effectiveness of a test which provides a free scope for socially desirable responses is drastically reduced.

Semantics

When responses to items in a test are worded with *often, sometimes, seldom, agree, slightly agree, strongly agree, disagree, slightly disagree, strongly disagree,* etc., these tend to create a semantic confusion among the examinees because what one subject calls 'seldom', another may describe as 'sometimes' and 'often'. Such confusion introduces error in the interpretation of the response category, thereby affecting the total score.

Evasiveness

The tendency to mark '?', 'uncertain', 'indifferent', etc., is known as evasiveness. Whenever the response alternatives are 'Yes-No' or '?', 'Agree-Disagree' or 'Indifferent', there is found a general tendency among the examinees to evade the two fixed categories (Yes-No) in favour of the doubtful category '?'. Somehow the examinees feel safe in marking this doubtful category. Cronbach (1946, 480) has reported that a score which was wholly based upon the number of doubtful categories had a reliability coefficient of 0.73. Guilford had also demonstrated a similar coefficient for such categories of responses. Obviously, such tendency to mark the doubtful categories, then, reveals a consistent habit on the part of the examinees, which must be researched.

Tendency to Work Speedily

Tendency to work for speed rather than accuracy is another common response set. Such a response set is commonly observed among tests of abilities or specific ability. Examinees having primarily ego-oriented motivation, go through the test of ability (in which ordinarily there is a reasonable time limit for giving responses to the items) rapidly without caring much for accuracy. Lengthy tests of temperament (say with 300 or more than 300 items) also encourage the examinees to respond speedily ignoring the true content of the items. Such a tendency distorts the score to the extent that it fails to reveal the true picture of the trait being measured and thus, the validity of the test is lowered.

Tendency to Guess when Uncertain

The tendency to guess when uncertain about the correct response is also a commonly observed response set. This occurs mostly in the situation where items are very difficult to answer and/or where items are given in a multiple-choice format or in matching form. Ability tests are more subjected to this type of response set than personality tests because in the former items are often written in either multiple-choice form or in matching form. Obviously, such a tendency contributes to the error variance of the score and, therefore, lowers the validity of the test.

Cautiousness

Cautiousness refers to a tendency to leave the items untouched or unanswered (particularly in the ability test or the special ability test), specially when the correct answer is in doubt. Such a response set is just the opposite of the tendency to guess when uncertain, and a large number of such responses, may adversely affect the test efficiency.

Extremeness

Extremeness refers to a tendency to mark the points lying only at the extremes of the rating scale rather than points towards the middle of the scale. If this tendency works consistently in giving answers to the items, it may also lower the efficiency of the test.

Researches on the above response sets particularly on social desirability, acquiescence and deviation set have passed through several stages. In the beginning, these response sets were regarded as the source of error variance which must be eliminated from the test scores. Later, these response sets were taken to be indicators of broad and durable personality characteristics worth measuring in their own right (Wiggins 1962). At this stage, they were commonly called as *response styles.*

IMPLICATIONS OF RESPONSE SETS

We have discussed the common types of response sets exhibited in answers to tests of ability, personality, interest, etc. These response sets have certain important implications for the test scores. These implications are briefly enumerated below.

1. Response sets may reduce the range of individual differences in test scores.
2. Response sets contribute to the error or irrelevant variance of the test scores. In other words, such sets tend to dilute the test with factors, which are not the part of the test content, and therefore, reduce the validity of the test.
3. Response sets in a test become more obvious and marked when items become difficult, ambiguous and unstructured. Hence, such items should be avoided.
4. Response sets are consistent. Researches on response sets reveal that response sets adopted by examinees are consistent from one administration to another administration of a test and from one test to another with some minor exceptions for a few examinees who may show a change over these administrations. As a consequence of this consistency, some of these response sets are now considered as general indicators of some traits of personality, which are worth measuring in their own right. Investigations by Wiggins (1962) and Jackson & Messick (1962) provide clear evidence of this. Response sets when considered as durable traits of personality are called *response styles,* though this distinction is not acceptable to all. Response sets being consistent over different administrations or from one test to another, tend to raise the reliability of the test score because consistency also contributes to the true variance of test scores. But such contribution is very small as compared to its contribution to the error variance and hence, such set should be controlled.

METHODS TO ELIMINATE RESPONSE SETS

While the different response sets may under certain conditions, where the test is measuring carefulness or other traits of personality that are similar to the response set itself, enhance the logical and empirical validity, their general impact is to reduce the validity of the test. Cronbach (1950) has called response sets the *enemy of validity.* As such, they must be eliminated from the test. In general, the following methods may be adopted in controlling the response sets:

Design of Test Items

The first preventive procedure in controlling response sets is to use such forms of test items, which may discourage the operation of response sets. Forms of the test items having fixed-response categories such as 'Yes-No', '?', 'True-False-Indifferent', 'Agree-Neutral- Disagree' and 'A', 'B', 'C', etc., have been found to encourage certain response sets like acquiescence, evasiveness, etc. To control these response sets the test technicians have recommended the use of 5-point, 7-point or 10-point scales of judgement in the place of the above three points. The scale requiring discrimination among 5 or 7 points tends to increase the reliability and the validity of the test scores by reducing the response set variance produced by the above two types of response sets, namely, acquiescence and evasiveness.

Investigators have also recommended the use of forced-choice items for controlling response sets such as faking-good or social desirability tendency and faking-bad or social undesirability tendencies. Forced-choice items require the examinees to choose between two (or more than two) terms or phrases which appear to them equally acceptable or unacceptable. The Edwards Personal Preference Schedule (EPPS) utilizes the forced-choice items with two terms whereas the Gordon Personal Inventory utilizes forced-choice items with more than two terms and the examinee is required to give two preferences in each item—one most descriptive of himself and another least descriptive of himself. But subsequent researches have shown that forced-choice items are not very effective in controlling the social desirability tendency of examinees. Besides, such items tend to introduce some administrative and technical difficulties, which further reduce their ability in controlling social desirability response sets. Another way of controlling social desirability as suggested by Meehl & Hathaway (1946) is to use an independent Social Desirability (SD) scale like the K scale of the Minnesota Multiphasic Personality Inventory (MMPI). Scores on this scale may be correlated with scores on the new test (whose social desirability is being controlled) to give the index of social desirability. Items yielding high and significant coefficient of correlation may be either reformulated or dropped. A third way of controlling the social desirability tendency as suggested by Wiener (1948) and Hanley (1956) is to select only those items which are of a relatively neutral type with respect to social desirability or undesirability. For this purpose Edwards (1957) has constructed a 9-point rating scale ranging from "extremely desirable" through neutral to "extremely undesirable". Judges may be requested to rate the items in terms of whether they consider each item desirable or undesirable in others. The median of the scale is, theoretically, 5, and the items having social desirability scale values ranging from 4 to 6 are considered as neutral with respect to social desirability or undesirability. Social desirability scale values above 6 are taken to be indicative of the faking-good tendency and below 4 to be indicative of the faking-bad. Edwards' 9-point scale has been presented in Table 8.1.

Table 8.1 Social desirability ratings scale

Ratings	Meaning of Ratings
1	Extremely undesirable
2	Strongly undesirable
3	Moderately undesirable
4	Mildly undesirable
5	Neutral
6	Mildly desirable
7	Moderately desirable
8	Strongly desirable
9	Extremely desirable

Adapting Test Difficulty

A response set increases with the rise in the difficulty level of items. As such, it is essential that the difficulty level of items be in accord with the ability of the examinees. Tests should be designed in a manner that if items prove to be difficult for a certain ability level of the examinee, another set of items which are relatively easier should be given to them. Tests of this type are very common in psychophysics.

Use of Correction Keys

Different types of the correction keys have also been utilized for controlling response bias. For example, in a famous personality inventory, namely, in the MMPI-2 where there are 567 affirmative statements to be answered by any of the three categories of responses—'True', 'False', and 'Cannot say'—there are four validating scales which are nothing but different types of correction keys. The four validity scores are the Question Score (?), the Lie Score (L), the Validity Score (F), and the Correction Score (K). Scores on these four scales, in general, indicate carelessness, faking-good, faking-bad, misunderstanding, evasiveness, etc. The first three scores—?, L and F—are usually used to evaluate the overall performance of the examinee on the inventory. If any of these scores goes beyond a certain specified value, the examinee's record of performance is treated as invalid. The K score is a kind of 'suppressor score'. A high K score indicates a deliberate attempt to fake-good and a low K score indicates self-criticism or a deliberate attempt to fake-bad. Since the K score is a 'suppressor score', it is computed as a correction score to be added to scores to get a final total score. The revised edition of MMPI, called MMPI-2, has three additional validity scales: Back F (F_b) scale, Variable Response Inconsistency (VRIN) scale and True Response Inconsistency (TRIN) scale. The F_b scale provides a check on validity and correction throughout the test. A high score on TRIN scale indicates tendency to respond true to most items (Yes-saying) and a very low score indicates to respond with false (No-saying). A high score on VRIN scale indicates that a large number of items were scored in a logically inconsistent manner.

Modifying the Directions

Directions may be written or modified to control some of the response biases. For example, if in the directions it is clearly mentioned that of the total true-false items about 50% items are actually false then this would automatically control the tendency to answer only the true options (acquiescence set of the examinees). Likewise, if directions are given to the effect that for every incorrect answer there is likely to be negative marking, it would automatically control the tendency to guess when uncertain. Similarly, evasiveness may be controlled by directing the examinees not to mark '?' or 'Uncertain' or 'Indifferent' as far as possible because that would be indicative of some undesirable tendency. Some investigators have demonstrated that the examinees should be directed to answer all items and as far as possible, a reasonable time should be given for the same so that accuracy may not be lost because of speed. One danger in directing the examinees to answer all items is that it may increase random error but in many cases this error is less damaging than constant errors produced by some response sets like tendency to guess, evasiveness, acquiescence, etc.

Use of Good and Appropriate Scoring Formulas

By developing good and appropriate formulas for the scoring of the correct items, incorrect items and unattempted items, the response set variance may be reduced to a great extent. According to Guilford (1954), the best solution is to obtain a factor analysis information regarding these different scores and accordingly, to weigh them.

Use of Easy Multiple-choice Items

Two-alternative items such as 'Yes-No', 'True-False', 'Agree-Disagree' encourage guessing on the part of the examinees. Moreover, the chances for guessing in these items are 50%. When multi-choice items are used in place of the two-alternative items, chances for guessing are automatically reduced and therefore, the response set variance is also reduced. Hence, the multiple-choice items have been regarded by many investigators to be the best ones.

One point that should be taken care of in constructing multiple-choice items is that they must match with the ability of the examinees. If necessary, different multiple-choice tests for the different levels of ability of the examinees may be constructed. Another point, if ignored, which is likely to produce response set variance, is regarding the position of the correct answers of each item. Correct answers should not be given at the same position in each item, rather their position or location should be randomly distributed. Mosier & Price (1945) have recommended that in order to control response set variance in multiple-choice items, both right and wrong answers should be randomized with respect to their position in the given set of alternative answers. For this purpose, they have provided a method.

☑ Review Questions

1. What do you mean by response bias? Discuss the important types of response sets often found in psychological tests.
2. Distinguish between faking-good and faking-bad. Point out the implications of response sets for psychological test.
3. Outline a programme for the methods of eliminating response sets.

◆

Part Two

PRINCIPLES OF MEASUREMENT

9

MEASUREMENT OF INTELLIGENCE, APTITUDE AND ACHIEVEMENT

CHAPTER PREVIEW

- Different Viewpoints Towards Intelligence
    - Intelligence As Whatever an Intelligence Test Measures
    - Intelligence As General Intellectual Capacity
    - Intelligence As a Combination of Groups of Traits or Factors
    - Intelligence As Structure of Intellect
    - Intelligence As Learning Ability
    - Intelligence As a Two-level Process
    - Intelligence As a Global Capacity
- Types of Intelligence Tests
    - Stanford–Binet Scale
    - Wechsler Scales
    - Downward Extension of the WAIS-III: the WISC-III and the WPPSI-R
    - Concluding Remarks on Wechsler Scales
    - A Comparison between Binet Scales and Wechsler Scales
    - The Kaufman Scales
    - Nonverbal Intelligence Scales
    - Some Indian Intelligence Tests
- Types of the Intelligence Test Scores
    - Mental Age
    - Intelligence Quotient
- Distinction between Aptitude Test and Achievement Test
- Types of Aptitude Tests and Achievement Tests
    - Aptitude Tests
    - Achievement Tests
    - Essay-type Tests Compared to Achievement Tests
    - Limitations of Achievement Tests
- Tests of Creativity

DIFFERENT VIEWPOINTS TOWARDS INTELLIGENCE

Tuckman (1975) in his 'Foreword' has said, "The modern history of testing is the history of testing for intelligence or mental ability." The term mental test was first used by J M Cattell in 1890. He was an American psychologist and studied with Wilhelm Wundt in Germany and Francis Galton in Great Britain. In fact, Cattell was a champion of what is now known as the 'brass instrument' approach to psychological testing because he demonstrated heavy reliance upon the use of brass instruments for measuring sensory thresholds and reaction times. This approach was based upon the assumption that keen sensory abilities were essential to intelligence. However, a student of Cattell demonstrated that sensory test results (that is, reaction time, naming of colours) bore little or no relationship to college grades. With such revelations, psychologists abandoned the use of reaction time and sensory measures as indices of intelligence. Today psychologists have made numerous attempts to define and measure intelligence. Despite their numerous attempts, the meaning of the term is not free from controversy, and there is not a single definition which is readily acceptable to all psychologists and educationists. In order to understand the meaning of this term, it is essential that the viewpoints of different experts be carefully examined. The experts, whose viewpoints are presented below, have defined intelligence (or its companion term 'general mental ability') from the different uses to which intelligence tests* have been put to.

Intelligence As Whatever an Intelligence Test Measures

Boring (1923) defined intelligence as *whatever the intelligence tests measure.* In other words, Boring's definition suggests that when we administer an intelligence test, we implicitly accept as intelligence whatever that test measures. This so-called definition was a favourite with those psychometricians who wanted to keep themselves away from theoretical disputes over intelligence. But this definition creates several difficulties. First, there are many tests for measuring intelligence. Which of these tests can be regarded as a standard measure of intelligence? The answer is not given by Boring's definiton. Second, many intelligence tests have a cultural bias and if Boring's definition is accepted, people would have to define intelligence differently in different cultures and there would be no common meaning of intelligence. Thus, it will create more confusion among the experts as well as people in general.

Intelligence As General Intellectual Capacity

Some psychologists have defined intelligence as a general intellectual capacity, which is made up of several discrete abilities. Binet (1919) has defined intelligence from this angle. According to him, intelligence is a general intellectual capacity which consists of the abilities: (i) to reason well with abstract materials (ii) to comprehend well (iii) to have a clear direction of thought (iv) to relate thinking with the attainment of a desirable end and (v) to be self-critical. Synthesizing the same, it can be said that he has defined intelligence as inventiveness that is dependent upon comprehension and marked by purposefulness and self-corrective judgement (Stanley & Hopkins 1978). Binet conceptualised intelligence as a single but a complex mental process which can be measured by various kinds of materials designed to measure the integrated mental processes. Binet did not really develop a theory of mental ability, rather he reinforced that mental ability developed with age, just as physical abilities did. He also felt that the intelligence was expressed in terms of various complex mental acts. That is the reason why he included a series of mental puzzles in his test, which he developed along with Simon. The second edition of the test was published in 1908 and he included tasks that were grouped into age levels, depending upon when the average child could do them. Binet died at a relatively young age in 1911, the year in which the third edition of his test was published. The same year William Stern, a German psychologist, introduced the term *intelligence quotient* for representing a child's standing relative to his or her age-mates. Spearman's (1927) viewpoints regarding intelligence are also placed under this category of general intellectual capacity. According to Spearman, intelligence is, by and large, a general intellectual capacity. He postulated the existence of a general factor (or *g* factor) and a specific factor (or *s* factor) underlying intelligence. All mental activities are mainly dependent upon and are an expression of this *g* factor. He defined the *g* factor as a mental energy that is required in all mental tasks and that is possessed by all individuals in varying degrees (because people differ in their mental activity). A positive correlation between two or more than two functions indicates the presence of the *g* factor. Another factor called the *s* factor, is a specific factor underlying a particular type of mental activity. Such a factor is unique to the activity itself. A low correlation between two or more than two functions indicates the presence of the *s* factor. Of the two, the *g* factor is more important because it is an important measure of intelligence. Spearman's view is known as the *two-factor theory of intelligence* which states that all mental activities have some general component (called the *g* factor) in common indicating general intellectual capacity, and each *g* factor has its own specific component (called the *s* factor), with the general one being the more important one. Spearman proposed that the aim of any intelligence test should be to measure the amount of *g* factor of each person because it provides the most important basis for predicting a person's behaviour in different situations. It is relatively of little use to measure the *s* factor because it is unique to a specific activity. Tests which measure abstract relations such as the Raven Progressive Matrices and the Cattell Culture Fair Intelligence Test, are also good measures of the *g* factor.

---

* Intelligence tests have been defined as tests to evaluate the general level of cognitive functions and hence, they are also known as tests of general cognitive ability. Intelligence tests are also called IQ tests, scholastic aptitude tests or academic aptitude tests.

The most difficult problem faced by Spearman's two-factor theory is the existence of group factors. In 1906, Spearman and his colleagues had noted that dissimilar tests could yield correlations higher than the values predicted from their respective *g* loadings. This finding automatically raised the probability that a group of diverse measures might share something common as a unitary ability other than *g*. For example, different tests might share a common unitary memorization factor that is really a halfway between *g* factor and the various *s* factors, which are unique to each test. Thus, the idea of the existence of group factors was incompatible with Spearman's two-factor theory.

Intelligence As a Combination of Groups of Traits or Factors

Intelligence has also been understood on the basis of the combination of groups of traits or factors. Thurstone (1938), L L Thurstone & T G Thurstone (1943) were the first persons to define intelligence on this basis and their theory is popularly known as the *group-factor theory* or *multiple-factor theory* of intelligence. They are of the opinion that intelligence is not an expression of the general factor, as postulated by Spearman, rather it is an expression of the combination of groups of traits or factors. Such factors are intermediate factors, not so universal and common as the *g* factor nor so truly specific as the *s* factor. According to the group-factor theory, some common mental activities have a primary factor on the basis of which they are distinguished from other mental activities. This primary factor gives the common mental abilities (or factors) a functional cohesiveness and then, they are said to constitute a group. Another group of common mental activities is said to have another primary factor and so on. In this way, there are a number of groups of mental abilities (or factors), each of which has its own primary factors. On the basis of extensive factor-analytic research, Thurstone & Thurstone (1943) postulated the following seven group factors which they designated as the *Primary Mental Abilties* or PMA: (These primary mental abilities do not incorporate the entire range of human abilties. They are specially from those found in abstract intelligence and in academic learning.) Factor analysis is a statistical procedure that identifies a group of tests, which correlate in such a way that they seem to share a common dimension.

1. The number factor (N) is the *ability to do numerical calculations rapidly and accurately*.
2. The verbal factor (V) is *found in tests involving verbal comprehension*.
3. The space factor (S) is *involved in any task in which the subject manipulates an imaginary object in space*.
4. The word fluency factor (W) is *involved whenever the subject is asked to think of isolated words at a rapid rate*.

5. The reasoning factor (R) is *found in tasks that require the subject to discover a rule or principle involved in series or groups of letters*.
6. The rote memory factor (M) *involves the ability to memorise quickly*.
7. Perceptual Speed (P) is *the ability to note visual details rapidly*.

It is obvious, therefore, that the factor-analytic research of Thurstone and Thurstone has revealed that intelligence comprises a number of aptitudes or factors or mental abilities.

Originally, each of these primary mental abilities was thought to be an independent one but later, the positive and significant correlations were demonstrated among these factors. The range of correlation was from 0.13 to 0.59. This meant that the primary factors or abilities were not the only factors in the operation of those mental activities which are assessed by various tests. They postulated that there must be some factors other than the primary factors or abilites to account for the positive correlations between those psychological tests which measure the above seven factors; they named these factors the *second-order general factors*. To this extent, they supported Spearman's viewpoint regarding some generality in the concept of intelligence. Thus, it may be concluded that intelligence is, to some extent, the expression of the combination of different factors (Thurstone) and, to some extent, the expression of some general factors (Spearman).

If we pay attention to the above interpretation, it is clear that Thurstone acknowledged the existence of *g* factor as a higher-order factor. By this time, Spearman had also accepted the existence of group factors that represented special abilities. Thus, it became apparent that difference between Thurstone and Spearman was largely a matter of focus or emphasis. Spearman continued to emphasize that *g* was the major determinant of correlations between test scores and ascribed a minor importance to group factors. Thurstone had reversed these priorities.

Vernon (1960) attempted to provide a rapprochement between Spearman's view and Thurstone's view by proposing a *hierarchical group factor theory*. According to Vernon, *g* was the single factor at the top of a hierarchy that included two major group factors labelled as verbal-educational (V : ed) and practical-mechanical-spatial-physical (K : M). Beneath these two major group factors were lying several minor group factors that resembled primary mental abilities of Thurstone. Specific factors were lying at the bottom of the hierarchy.

In 1966, Raymond Cattell and John Horn developed their own viewpoints about intelligence and they postulated *fluid intelligence* and *crystallized intelligence* as the two most important factors of intelligence. Fluid intelligence is largely nonverbal and culture-reduced form of mental ability. Therefore, it may be understood as problem-solving and information-processing ability that is by and large independent of experience. Crystallized intelligence represents what one has learnt through the investment of fluid intelligence in the culture (for example, learning trigonometry in schools). Thus, crystallized intelligence is culturally dependent and is primarily used for those tasks that require a learned response. Since crystallized intelligence arises when fluid intelligence is applied to our life experiences derived from the cultural background, it is expected that these two kinds of intelligence will be correlated. It has been commonly found that the tests of fluid intelligence and crystallized intelligence correlate moderately ($r = 0.50$) (Gregory 2006). Because both of these mental abilities are general and broad based, they were given the symbols G_f (general fluid ability) and G_c (general crystallized ability). In subsequent years, Horn and his colleagues expanded the number of abilities associated with the theory to nine factors which are as follows:

1. G_f (fluid intelligence) such as ability to reason in novel situation
2. G_c (crystallized intelligence) such as depth of general knowledge
3. G_q (quantitative ability) such as ability to comprehend and understand numerical concepts
4. G_a (auditory processing) such as ability to discriminate sounds and detect sound patterns
5. G_v (Visualization processing) such as ability to see spatial relationships and patterns

6. G_s (Processing speed) such as ability to take quick decisions and maintain attention
7. G_{sm} (Short-term memory) such as ability to use certain information over a short period of time
8. G_{lr} (Long-term memory) such as ability to transfer material to permanent storage and retrieve it later on
9. CDS (Correct decision speed) such as ability to arrive at correct decision soon

John Carroll (1993) undertook to reanalyze the data from more than 400 factor analytic studies of abilities for finding out some common themes. The result was a hierarchical model similar to that of Vernon's model. One feature of this model was three-stratum explanation of intelligence. The general intellignece or *g* was at the top (stratum III) of the model; a small number of general-ability factors similar to those of Cattell–Horn model was at the second level (stratum II); and a large number of specific factors describing the abilities required for performance on narrower test classification at the lowest level (stratum I). Recently, Carroll and Horn have expressed the view that there lies much similarity between $G_f \cdot G_c$ theory and three-stratum theory and therefore, they have named the combined theory as *Cattell-Horn-Carroll (CHC) theory of intelligence.*

Intelligence As Structure of Intellect

Intelligence has also been defined on the basis of structural considerations of discrete factors. On the basis of his factor-analytic research of nearly 20 years, Guilford (1967) has proposed a three-dimensional boxlike model which he calls the structure of intellect model or SI model. The model has tried to simplify the picture of intellectual trait relationships by organizing the traits along three dimensions viz., *contents, operations and products*. Each of these aspects of intelligence was analyzed and separated into subcategories: five for operations, six for products and five for content, making a cube of $5 \times 6 \times 5 = 150$ abilities. Shortly before his death, Guilford (1988) expanded his theory from 150 to 180 abilities when he divided the operation of memory into two categories—memory recording and memory retention (Wood & Wood 1996). Thus 6 abilities for operation, 5 for content and 6 for products constituted $6 \times 5 \times 6 = 180$ abilities. A description of these three categories and subcategories is given below.

OPERATIONS: It refers to the basic intellectual processes of thinking used by persons. It has the following five subcategories:

C Cognition: It includes discovery, rediscovery, recognition of information or some understanding.
M Memory Recording: It includes a person's ability to readily encode the information.
Memory Retention: It includes a person's ability to retain the encoded information.
D Divergent Production: It refers to the ability to search for multiple, creative or novel solutions to a problem.
E Evaluation: It means placing a value judgement on knowledge and thought.
N Convergent Production: It includes the ability to search for a correct solution to a problem.

CONTENT: It refers to a type of content or material on which operations are performed, that is, terms in which the persons think. It has the following five subcategories:

V Visual: It includes concrete visual materials.
A Auditory: It includes nature and characteristics of sound perceived.
S Symbolic: It includes letters, words, digits and other similar signs.
M Semantic: It includes verbal meanings or ideas.
B Behavioural: It includes knowledge regarding other persons.

PRODUCTS: It refers to the results of performing operations on contents, that is, the form of thought produced by individuals. It has the following six subcategories:

- U Units: It refers to the production of a single word, definition or isolated bit of information.
- C Classes: It refers to the production of a concept.
- R Relations: It refers to the production of any form of relationship such as an analogy, an opposite or other similar ones.
- S Systems: It refers to the production of an internally consistent set of classifications of various forms.
- T Transformation: It refers to the production of changes in meaning, organization or some other arrangements.
- I Implications: It refers to the production of such information as is beyond the data presented.

In the Guilford model each cell in the cube should be reasonable, and over 100 such factors have been identified so far. In this model each cell in the cube is identified by a three-letter code (cf. Figure 9.1) that refers to an *operation*, a *content* and a *product* in that order. In Figure 9.1, each of the four corner cells has been separated and shown by a separate cube to facilitate understanding. For example, CVU is one such cube on the upper left-hand side and it is *cognition of visual units*. Likewise, three other cubes are EBU (Evaluation of Behavioural Units), EBI (Evaluation of Behavioural Implications) and CBI (Cognition of Behavioural Implication). Let us illustrate the use of the model with an example. Suppose a little child is asked to locate the day of the week on a particular date in a calendar. This task obviously involves operations like Memory (M), Cognition (C) and Convergent thinking (N). In carrying out these operations he will be making use of the contents like Semantics (M), that is, reading and understanding the printed words and figures that may indicate the date and day of the particular month. By carrying out these mental operations with the help of contents, the child will arrive at the products. The day of the week to which the date in question refers represents the factor known as Relations (R).

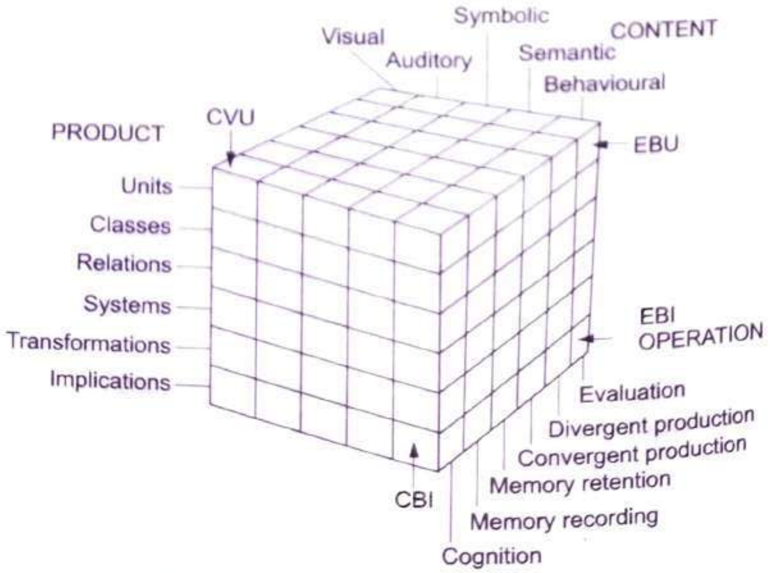

Fig. 9.1 Guilford's structure of intellect

Although one test has been directly designed on this Guilford's model, namely, the structure of Intellect learning Abilities test (M Meeker, R Meeker & Roid 1985) the model itself has little impact upon development and application of tests for general use.

Intelligence As Learning Ability

Intelligence has also been defined as the ability to learn. In this sense, an individual's intelligence is a matter of the degree to which he or she is educable. If a person is able to learn something readily and quickly, he is said to be an intelligent person. Piaget (1950, 1952) defined intelligence on the basis of assimilation and accommodation, which together determine a person's ability to learn. *Assimilation* denotes changes occurring in what has entered into the mind from the external world in such a way that they fit into a frame of reference, and *accommodation* denotes changes occurring in the internal structure of a person as a function of new experiences. If this definition of intelligence is taken as a base, major changes in the modern approach to intelligence testing would be required because the modern approach emphasizes most what a child has learned whereas this definition emphasizes what a child can learn as a result of proper instruction. Feuerstein (1975) has developed a procedure in which a child's learning ability is measured after he is properly instructed for sometime.

Intelligence As a Two-level Process

Intelligence has also been defined on the bais of a combination of two levels of processes. At one level is found *associative intelligence* and at another *abstract intelligence*. Jensen (1968) has defined intelligence from this viewpoint. Associative intelligence includes those kinds of tests that depend upon memory as well as simple verbal associations. Such intelligence may include factors like verbal associations, memory for temporal sequences, spatial positions, and stimulus-response learning. Abstract intelligence, on the other hand, includes factors like concept learning, thinking, problem-solving skills, multiple discrimination, principle learning, etc. (Gagne 1965; Freeman 1962). Most intelligence tests are a combination of some factors from associative intelligence and some from abstract intelligence. Associative intelligence relates to biological maturation and shows little variation with different social classes and races. Abstract or cognitive intelligence is dependent upon education and culture and is, therefore, responsible for the observed differences in the abilities among different social classes and races.

Intelligence As a Global Capacity

Intelligence has been defined as a global capacity or a composite of several intellectual skills. Wechsler (1944) and Stoddard (1943) have defined intelligence as such. According to Wechsler, intelligence is "the aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment". One of the salient features of this definition is that intelligence is displayed by the behaviour of the individual as a whole and that intelligent behaviour is goal-directed and helps in making effective adjustment in the given environment. Wechsler, thus, has included the concept of "drive" and "incentive" which are implied in his statements "to act purposefully" and "to deal effectively". But his viewpoint has been criticised by many experts who are of the view that drive and incentives are nonintellectual traits of personality and if they are included in a test of mental ability, it will create more confusion (Freeman 1962). Similarly, Stoddard has defined intelligence as "the ability to undertake activities that are characterized by (i) difficulty (ii) complexity (iii) abstractness (iv) economy (v) adaptiveness to goal (vi) social value and (vii) the emergence of originals and to maintain such activities under conditions that demand a concentration of energy and resistance to emotional forces." Of these seven characteristics three skills, namely, economy, social value and the emergence of originals, deserve special attention because these attributes have not been discussed so far in any approach. *Economy* indicates the rate at which a mental task is done or a problem is solved. If A solves a problem sooner than B, A would be regarded as more able than B and would be said to have the characteristic of economy. The term 'social value' indicates whether or not the mental tasks performed by a person are in accord with socially desirable and acceptable norms. If they are, the task is said to have the characteristic of social value. The inclusion of social value in the definition of intelligence by Stoddard has aroused a lot of controversy and confusion. Psychologists, in general, are of the opinion that 'social value' has subjective element and therefore, it is difficult to judge whether or not a given mental operation is socially desirable or undesirable. As such, such an attribute should not be included in any scientific definition of intelligence nor in any intelligence test. The *emergence of originals* refers to a person's ability to discover something new and different, that is, it indicates the ability of creativity and originality. A few examples of such ability are to discover a new relationship and a new scientific principle from the given data and to discover a new fact or theory regarding certain psychological variables or social facts. Stoddard's definition has, however, been criticised because of this inclusion of the two conditions of intelligent behaviour, that is, concentration of energy and resistance to emotional forces. It is said that motivational and emotional forces are nonintellectual traits and should not be treated as aspects of mental abilities nor should they be included in any intelligence test. If this is done, it would be to confuse and invalidate measures of the intelligence.

Intelligence As a Process of Integrating Physiology and Information Processing

Das, Naglieri and Kirby (1994) explained intelligence as a process of integrating physiology and information processing. Their model is known as PASS theory (or model) of intelligence. On the basis of neurophysiological model of brain proposed by A R Luria, these experts have explained intellectual performance into four basic processes: *Planning* (P), *Attention* (A), *Simultaneous scanning* (S) and *Successive processing* (S), popularly synthesized as PASS model. The model relates these four processes to specific neurological structures or areas, which Luria called functional units (Das & Naglieri 2001). Luria described the cognitive processes within the framework of three functional units. The function of the first unit is cortical arousal and attention; the second unit codes information using simultaneous and successive processes; and third unit is responsible for providing, self-monitoring and structuring of cognitive activities. This work of Luria on functional aspects of the brain structures became the basis of PASS model.

The most important and basic function is attention. A person must attend to a stimulus in order to process the information. Any inability to focus attention on the stimuli is seen as one primary reason of the poor intellectual performance. Once the attention is focused on the stimuli, the resulting information may require either simultaneous focusing or successive focusing. Information that is subject to simultaneous scanning is said to be surveyable because elements are interrelated and accessible. Simultaneous processing is involved in language comprehension and other such tasks that require the perception of whole stimulus situation at once. Successive processing is required whenever the elements of the task have to be completed in a particular order. Tasks like serial recalling of stimuli and performing some skilled movements require successive processing.

The planning part of the model is involved in deciding where to focus attention and what type of processing is required by the given task. This portion is very similar to Sternberg's metacomponents. This planning process also monitors the success of problem solving and if needed, modifies the approach until a solution is arrived at. These four processes, in fact, operate within an existing knowledge base, which is thought to be the cumulative result of an individual's experiences stored in the memory. Since all four processes operate within the context of knowledge base, the base of information tends to influence all cognitive processes (Das, Naglieri & Kirby 1994).

In conclusion, it can be said that the term 'intelligence' is difficult to define in a language which is acceptable to all. However, intelligence can be understood as a general set of mental traits, which is often reported as *mental abilities*. Often, intelligence tests are confused with tests of creativity which measure a similar but different aspect of the mental process. Tests of *mental ability* or intelligence tests measure convergent thinking whereas tests of creativity measure primarily divergent thinking. Convergent thinking is one where the individual tends to reduce a large number of information to one or two possible correct answers whereas in divergent thinking, the individual tends to expand from a limited amount of information to many correct answers. The Torrance tests of creativity and Remote Associate tests are two common examples of test of creativity which have been discussed a little later in this chapter.

TYPES OF INTELLIGENCE TESTS

Intelligence tests have been classified from different angles. From the point of view of their administration, they are divided into two convenient categories—individual tests and group tests. An *individual intelligence test*, as its name implies, is one which can be administered to one person at a time. The first individual intelligence test was the Binet–Simon scale. A *group intelligence test* is one which can be administered to more than one person at a time, that is, it can be administered to a group. The first group intelligence test was developed when the necessity for mass testing was realised during World War I. The Army Alpha test and the Army Beta test were the first group tests, the former being a verbal one and the latter being a nonlanguage one. Besides the above simple comparison, the individual test and the group test may be compared on the following points so that the two may be recognized as two distinct tests:

1. The individual test requires a highly skilled and experienced test administrator. He must have a specialized knowledge of testing procedures and test administrations. The group test, on the other hand, being mostly self-explanatory in nature, does not require a well-trained and skilled administrator. Persons having a moderate experience with testing procedures can do well in the case of administration of the group test.

2. Individual tests are mostly applied in clinical settings whereas group tests are mostly applied in educational settings, industries, the civil services and military services. A person who is discouraged, withdrawn and has a strong sense of inferiority and guilt cannot be tested properly in a group. However, in an individual session with an experienced adminstrator he can be readily diagnosed with the help of the appropriate test.

3. Individual tests are most suitable for very young children whereas group tests are most suitable for adolescents and adults. With very young children, that is, children of preschool ages, only individual tests can be administered. No group test can be successfully administered with them. There are two reasons for this. First, the children of preschool age are distracted very easily. If they are asked to sit in a group, they will be even more distracted. The only solution is to test them individually where the test administrator can have a better control over the distraction of a child. Second, such children cannot be motivated to do well in group tests. The administrator cannot encourage and motivate each child of a group to the extent he deserves. In an individual session, the same task becomes easier for him.

4. Individual tests are usually more difficult to construct than group tests. The construction of an individual test is a time-consuming process and an expensive job. Preparing test items for individual tests, standardizing those items on a suitable representative sample, and then finding the proper instructions, time limits, method of scoring are definitely arduous, noneconomical and time-consuming processes. This is specially so because the entire process of test construction is to be carried out in an individual session. It is, therefore, advisable that individual test construction should not be undertaken unless the test constructor is ready to expend a considerable amount of money and time in test construction.

5. In group tests, the administrator has a very little opportunity to establish rapport and motivate the examinees and this is especially true when the group is large. Not only this, in the group test the administrator cannot readily recognize those examinees who are being influenced by some transient conditions like headache, fatigue, worry, etc. In individual tests, the

administrator has a better chance to detect the influence of transient conditions upon the examinees, and he is also better able to establish a rapport and obtain cooperation from the examinee.

6. The norms of group tests are more dependable and established than the norms of individual tests. Because of ease in data collection, the norms of group tests can be based upon a very large sample but in case of the individual tests, norms cannot be based upon a very large sample because of the difficulty and the time-consuming nature of data collection.

Intelligence tests have also been classified on the basis of the nature of items used in tests. The verbal test (or paper-and-pencil test) and the performance test are the natural outcome of such a classification. Strictly speaking, a *verbal test* is one in which the instructions and items are reproduced usually through the written language before the examinees. It automatically means, then, that for a verbal test the examinee must be literate because he will have to use a pencil and paper for answering the items. Such tests are also called paper-and-pencil tests of intelligence by laymen. All verbal tests by their nature are group tests. The Army Alpha test and the Army General Classification test are examples of verbal tests. The Mohsin General Intelligence test, the Joshi General Mental Ability test and the Jalota's Group General Mental Ability tests are some examples of verbal tests developed by Indian psychologists.

A performance test is one where language is used only to impart instruction and items are manipulative in nature. The examinee is required to answer by manipulating the given test materials. No written language is needed to answer the items. Thus a performance test is not a paper-and-pencil test. It can be administered to literates as well as illiterates, speech defectives, preschool children who cannot use language, and on the culturally deprived, etc. All performance tests by their nature are individual tests. Examples of performance tests are the Army Beta test, the Pinter–Patterson scale, the Cornell–Coxe scale, the Arthur Point scale, the Pass-along test, and the Kohs Block Design test. In addition to the verbal test and the performance test, there is another important class of intelligence test commonly known as the nonverbal test or the *nonlanguage test* or the culure-free test. By definition, a nonverbal test or a nonlanguage test is one where no language is used at all either in instruction or in construction of items. Usually, the instruction is given through pantomime, gestures, blackboard demonstrations and charts and the items are neither verbal nor manipulative types. Test items are usually of figured relation type where the examinee is to discover, without the actual manipulation of objects, the relationship between various figures and designs. In a nonlanguage test, the use of paper and pencil is involved. This makes it similar to a verbal test and distinct from a performance test. The examinee uses the paper and pencil only to underline or cross out the items, which do not require any ability to read or write. The use of paper and pencil in a nonlanguage test is limited to making only nonlinguistic remarks. Nonlanguage tests are usually group tests and are often given when the examinees are incapable of being tested with the usual tests of intelligence. Nonlanguage tests are most suitable for subnormals, physically handicapped, foreign-language persons, etc. One of the biggest advantages of a nonlanguage test is that it can be administered to persons belonging to different cultures. Hence, it is also sometimes known as *culture-free or culture-fair or cross-cultural or culture-reduced* test of intelligence. The first nonlanguage test was developed during World War I for illiterate soldiers. The name of the test was the Army Examination Beta. Some of the other tests which are regarded as culture-free tests are the Leiter International Performance scale, the Culture Fair intelligence test by Cattell, and the Raven Progressive Matrices test and the Goodenough Draw-a-Man test, now known as Goodenough-Harris Drawing test. But strictly speaking, these tests are not free from cultural influences. As a matter of fact, no one has fully succeeded in constructing a culture-free test in the true sense of the term. The only thing that can be said in favour of these tests is that they share the maximum experiences common to a large number of cultures. Hence, the more appropriate name for them would be *Culture-Common* or *Culture-General* tests rather than culture-free tests.

A detailed description of some very important individual intelligence scales is presented below.

Stanford–Binet Scale

A very important step in the development of the individual intelligence was taken by Alfred Binet (1857–1911) who established the first psychological laboratory in 1889 in France. He was asked by the Government of France to devise a means for diagnosing slow learners and mentally retarded children in Paris schools. Working with Simon, a physician at the asylum at Saint-Yon, Binet developed an intelligence test for this purpose. The test is known as the *Binet–Simon scale*, and was first published in 1905. The Binet–Simon scale consisted of 30 items, which were arranged in ascending order of difficulty. The scale was a crude measure of intelligence of school-going children. In 1908, Binet and Simon revised the scale in order to remove some of its defects. The 1908 Binet–Simon scale was the first age scale and it created considerable interest among psychologists working in different countries like Germany, England, Belgium, Switzerland, Italy, and the United States. As a result, many suggestions and criticisms were made and in the light of these suggestions and criticisms, Binet and Simon further revised the scale in 1911 in which the age range was extended from three years to the adult level. This was the last revision by Binet in his lifetime.

In the United States of America, several adaptations and translations of the Binet–Simon scale were developed. The most important of these adaptations, which are today very popular, were done at Stanford University. There were, however, four other revisions in the United States of America. In 1908, H H Goddard published an English translation of the 1905 Binet–Simon scale, and in 1911 he published a revised version of the 1908 Binet–Simon scale. Other revisions in the USA were made by F Kuhlman in 1912, and 1939; R Yerkes in 1915 and 1923; and J P Herring in 1922. But these four revisions are remembered today only for their historical interest. The first important American revision of the Binet–Simon scale, which gained prominence, was done by Terman and his associates at Stanford University in 1916. In this revision, the Binet–Simon scale was given almost a new look. More than one third of the items were new and the entire scale was restandardized on an American sample of 1400 in which 1000 were children and 400 were adults. The most important aspect of this revision was the concept of IQ, which was for the first time introduced in a psychological test. In 1937, Terman along with Merill revised the 1916 Stanford–Binet test which is commonly referred to as the 1937 Binet. This revision comprised two equivalent forms—L and M. The 1937 Binet was better standardized and better validated than the 1916 revision and it also extended the upper and lower range. The third revision of the Stanford–Binet scale was done in 1960. This revision consists of a single form in which the test items from both L and M forms were retained. Hence, the 1960 Binet is known as the L–M form. Generally, the 1960 revision has retained the important characteristics of the 1937 revision. However, it has dropped the unsatisfactory subsets of the 1937 revision, and has relocated a few items. The 1960 revision has a range of 2 years of mental age scores to 22 years and 11 months of mental age score. The single L–M form is now available with norms on data as recent as 1972. According to Sattler & Theye (1967), this form measures abilities from seven categories: language, reasoning, memory, social intelligence, conceptual, numerical reasoning and visual-motor. Test items are in the form of words, objects and pictures and the response given by the testees are in the form of drawing, calculating, writing, and speaking. About an hour is required in its administration. In this revision, the intelligence is expressed in terms of standard score of the deviation IQ.

The next stage was the 1972 restandardization of Form L–M. The 1960s revision had failed to include a new *normative sample* or restandardization. However, by 1972, a new standardization group consisting of a representative sample of 2100 children (about 100 at each Stanford–Binet age level) had been obtained for use with the 1960 revision (Thorndike 1973). One basic feature of the 1972 norms was that unlike all previous norms, it included nonwhites.

The new edition, that is, fourth edition of the Stanford–Binet (SB4) appeared in 1986. This edition, in fact, represents its most extensive revision (Delaney & Hopkins 1987; Thorndike, Hagen & Sattler 1986). This revision, on the one hand, retained the chief advantages of the earlier editions as individually administered tests and, on the other hand, it also reflected the intervening developments in both theoretical conceptualizations of intellectual functions and methodology of test construction.

This modern Binet scale incorporates the *gf-gc* theory of intelligence (Horn & Noll 1997). According to this theory, there are two basic types of intelligence: fluid (*gf*) and crystallized (*gc*). Fluid intelligence can best be thought of as those abilities that allow the person to reason, think and acquire new knowledge. Crystallized intelligence, on the other hand, displays the knowledge and understanding that the person has acquired. The modern Binet scale is based upon three levels of hierarchical model as shown in Figure 9.2.

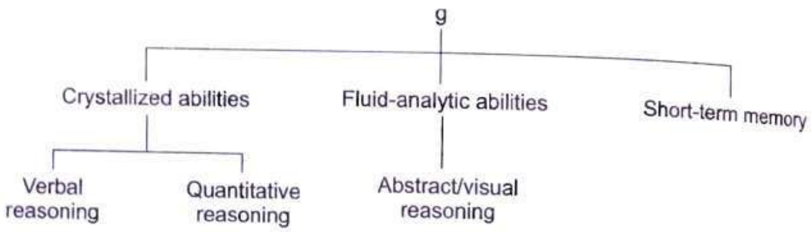

Fig. 9.2 Three-level model of Binet's modern scale

As it is obvious from Figure 9.2, at the top of model is the g-factor (or general intelligence) which, in fact, reflects the common variability of all tasks. At the next level of the model are three group factors—Crystallized abilities reflecting learning, Fluid-analytic abilities reflecting the original potential that a person uses to acquire crystallized abilities (Horn 1994; Horn & Cattell 1966) and Short-term memory reflecting a persons's capabilities to retain information briefly after a single brief presentation. Crystallized abilities are further divided into two subcategories—verbal reasoning and quantitative reasoning. Fluid-analysis reasoning is reflected by abstract/visual reasoning. The component of verbal reasoning is assessed by four tests, quantitative reasoning is assessed by three tests, abstract/visual reasoning is assessed by four tests and short-term memory is assessed by four tests. In this way, Binet's modern scale consists of 15 subtests as shown in Figure 9.3.

Binet's modern scale permits the calculation of specific scores for each of the 15 tests. Thus in addition to an overall score that reflects g, the test users can obtain scores related to each specific content area. In order to avoid having wider content unevenly distributed across age groups, the authors dropped the age scale format in this version. Now in place of age scale, items having the same content were placed together into any one of 15 separate tests. For example, all matrix items were placed together in one test and all vocabulary items were placed together in another test. In addition, each of the specific 15 tests can be grouped into one of four content areas. This scale measures intelligence in individuals aged 2 years through adult. Due to this, all items in a given test will not be appropriate for all subjects. Therefore, in administering 1986 Binet, one has to determine from where to begin each specific test. This is done by using the subject's individual score on the vocabulary test in conjunction with his or her chronological age. Thus the test format is *adaptive*.

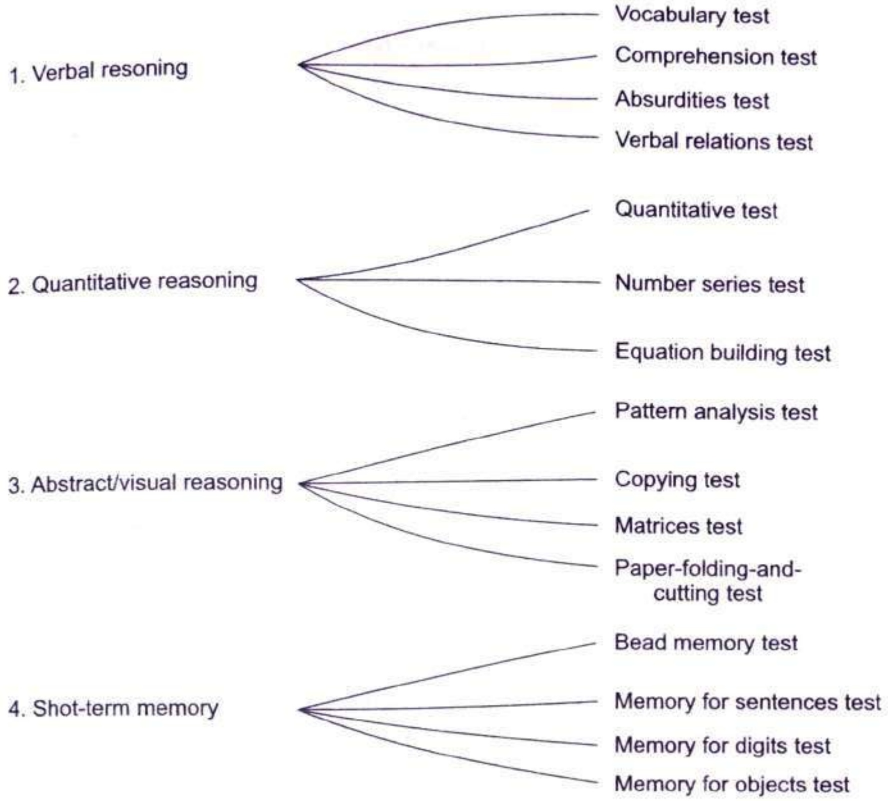
1. Verbal resoning
Vocabulary test
Comprehension test
Absurdities test
Verbal relations test
2. Quantitative reasoning
Quantitative test
Number series test
Equation building test
3. Abstract/visual reasoning
Pattern analysis test
Copying test
Matrices test
Paper-folding-and-cutting test
4. Shot-term memory
Bead memory test
Memory for sentences test
Memory for digits test
Memory for objects test

Fig. 9.3 Four major areas of Binet's modern scale and the related tests

The test users must also establish the *basal age* for each test, that is, the lowest level point where two consecutive items of approximately equal difficulty are passed. In this way testing proceeds within each test until a ceiling is reached, that is, point at which at least three out of four items are missed. The concept of age differentiation has been retained, that is, items are presented in order of increasing difficulty according to their ability to discriminate different age groups. Thus the age differentiation refers to the fact that with increasing age, children develop their abilities.

Binet's modern scale retains the use of standard scores. Accordingly, raw scores on each of 15 tests can be converted to a *standard age score* with a mean of 50 and a standard deviation of 8. The subgrouping of individual tests into four content areas permits one to calculate four area-content scores, each with a mean of 100 and standard deviation of 16.

Like any other individually administered test, the 1986 revision of Stanford-Binet is a good measure of general intelligence. Its norm was developed on adequate standardization sample. Administration and scoring instructions are clear and the test discriminates well among normal, gifted and neurologically impaired samples.

However, there are some limitations. *First,* there exists disagreement about the factor structure of the instrument and this ultimately raises a question about the appropriate interpretation. It appears that the factor structure shifts from a two-factor to a four-factor organization as the sample advances in age. *Second,* it is often a difficult task to establish a basal level with younger, intellectually impaired children. *Finally,* the administration of the test is somewhat more difficult than other tests since the examiner has to score each item as the test is administered.

The latest edition is the fifth edition of Stanford–Binet test (SB5) and it was introduced in 2003 by Roid. The items of this new scale were reviewed for gender, ethnic, cultural, regional and socio-economic biases. The fifth edition is organized in a little different way from the fourth edition. It produces a Full scale IQ, Verbal IQ and Nonverbal IQ and five composite Indices. Another remarkable change in the fifth edition is that the standard deviation for the various scores have been set at 15 as opposed to 16 which was the tradition with earlier versions of the test.

The organization of SB5 test has been guided by the principle that each of the five factors of intelligence may be assessed by two distinct domains: *Verbal* and *Nonverbal*. The five factors are *fluid reasoning, knowledge, quantitative reasoning, visual-spatial processing* and *working memory*. When these five factors are crossed over two domains (verbal and nonverbal), the result is an instrument with ten subtests like *nonverbal fluid reasoning, verbal fluid reasoning, nonverbal knowledge, verbal knowledge, nonverbal quantitative reasoning, verbal quantitative reasoning, nonverbal visual-spatial processing, verbal visual-spatial processing, nonverbal working memory* and *verbal working memory*. Thus SB5 provides ten subtest scores (mean of 10 and SD of 3), three IQ scores (full scale IQ, verbal scale IQ and nonverbal scale IQ) as well as five factor scores (fluid reasoning, knowledge, quantitative reasoning, visual-spatial processing and working memory). The IQ and factor scores have norms based upon a mean of 100 and SD of 15.

The basic features of SB5 may be enlisted as under:

1. The developers of SB5 have screened items for fairness based on variables like gender, race, ethnicity, disability and religious tradition (Christian, Jewish, Muslim, Hindu and Buddhist backgrounds). In fact, this is for the first time in the history of intelligence testing that religious tradition has been considered.

2. SB5 is suitable for assessing intelligence for children aged 2 through adults aged 85 and older.

3. SB5 uses a *routing procedure* for estimating the general cognitive ability of the examinee before proceeding towards the remainder of the test. The aim of routing procedure is to identify the appropriate starting points for later subtests. The routing items belong to both verbal and nonverbal domains. These items also provide what is called *Abbreviated IQ* which are sometimes used for screening purposes.

4. The working memory factor, which consists of both verbal and nonverbal items, helps in assessing and understanding children with attention-deficit/hyperactivity disorder.

In sum, it can be said that SB5 is a very promising intelligence test that is useful at both ends of the cognitive spectrum—the very young and the very gifted person. It is likely to be a mainstay of individual intelligence testing. In Table 9.1, some important milestones in the development of SB5 and its predecessors have been presented for better sequential understanding.

Table 9.1 Milestones in the development of SB5 and its predecessors

Year	Test/Authors	Basic feature
1905	Binet and Simon	Simple 30-item test arranged in ascending order of difficulty
1908	Binet and Simon	Mental age concept introduced
1911	Binet and Simon	Extended to cover adults
1916	Stanford–Binet (Terman and Merrill)	The concept of IQ introduced
1937	Stanford–Binet-2 (Terman and Merrill)	First use of parallel forms (L and M

1960	Stanford–Binet-3 (Terman and Merrill)	Modern item analysis methods used
1972	Stanford–Binet-3 (Thorndike)	SB-3 restanderdized on 2,100 persons
1986	Stanford–Binet-4 (Thorndike, Hagen and Sattler)	Complete restructuring into 15 subtests
2003	Stanford–Binet-5 (Roid)	Five factors of Intelligence introduced

Wechsler Scales

In order to remove the difficulties and shortcomings of the Stanford–Binet scale, particularly the difficulties regarding the measurement of adult intelligence, David Wechsler, working at the Bellevue Hospital at New York city, developed a new scale for measuring adult intelligence in 1939. The scale was known as the *Wechsler–Bellevue scale* to be used with children above 10 and adults. The Wechsler–Bellevue scale was developed in two equivalent forms—Form I and Form II. Each form consisted of 10 subtests of which five were verbal tests and five were nonverbal or performance tests. In 1955 the Wechsler–Bellevue scale was revised and renamed as the *Wechsler Adult Intelligence Scale* or WAIS, which consisted of 11 subtests, of which six were verbal scales and five were nonverbal or performance scales. It was called WAIS-R after a revision done in 1981 and WAIS-III after revision in 1997.

The WAIS-III consists of seven verbal subtests and seven performance subtests. The seven standard verbal subtests and the functions they intend to measure are presented in Table 9.2

Table 9.2 WAIS III: Verbal subtests and their functions

Names of the verbal subtests	Major function measured
1. Vocabulary test	Vocabulary level
2. Similarities test	Abstract thinking
3. Arithmetic test	Concentration
4. Digit span test	Immediate memory, anxiety
5. Information test	Range of knowledge
6. Comprehension	Judgement
7. Letter-Number sequencing test	Freedom from distractiblity

The seven standard performance subtests and the functions they intend to measure are presented in Table 9.3.

Table 9.3 WAIS III : Performance subtests and their functions

Names of the performance subtests	Major functions measured
1. Picture-completion test	Alertness to details
2. Digit Symbol-coding test	Visual-motor functioning
3. Block design test	Nonverbal reasoning
4. Matrix reasoning test	Inductive reasoning
5. Picture arrangement test	Planning ability
6. Symbol search test	Information-processing speed
7. Object-Assembly test	Analysis of part-whole relationship

The seven verbal subtests together make up the *verbal scale* and the seven performance subtests together make up the *performance scale*. Each subtest produces a raw score, that is, a total number of points. For making a comparison of raw scores on the individual subtests, the raw scores can be converted to *standard* or *scaled scores* with a mean of 10 and standard deviation of 3. Two sets of norms have been derived for this conversion: *age-adjusted norms* and *reference-group norms*. Age-adjusted norms were based upon a normalized cumulative frequency distribution of raw scores for each age group, and reference-group norms were based on the performance of participants in the standardization sample between ages 20 and 34 (Tussky, Zhu & Ledbetter 1997). To obtain the verbal IQ (VIQ), the age-corrected scores of the verbal subtests are added together. The verbal IQ is a deviation IQ with a mean of 100 and a standard deviation of 15.

Like verbal subtests, the raw scores for each of the seven performance subtests are converted to scaled scores with a mean of 10 and standard deviation of 3. The performance IQ (PIQ) is obtained by summing the age-corrected scaled scores on the performance subtests. Like verbal IQ, performance IQ is also deviation IQ with a mean of 100 and standard deviation of 15.

In addition to verbal IQ and performance IQ, there is a provision for Full scale IQ (FSIQ). It is obtained by adding the age-corrected scaled scores of verbal subtests with the performance subtests and comparing the subject to the standardization sample.

In addition to the verbal IQ, performance IQ and full scale IQ, the WAIS-III provides opportunity for computing index scores. In this scale, there are four index scores: *verbal comprehension index, perceptual organization index, working memory index* and *processing speed index*. A schematic overview of WAIS-III has been presented in Table 9.4. Four index scores tend to assess different aspects of human intelligence.

The verbal comprehension index measures crystallized intelligence. The perceptual organization index measures fluid intelligence. The working memory index measures ability to hold manipulations and information in our mind. In fact, the concept of working memory refers to the information that the person actively holds in mind, in contrast to our stored knowledge or long-term memory. Processing speed index measures how quickly the mind of the person works. For example, one person may solve the given problem in 10 seconds whereas another person may solve the same problem in 50 seconds.

Table 9.4 A Schematic overview of WAIS-III index scores

1. Verbal Comprehension Index	Vocabulary test
	Similarities test
	Information test
2. Perceptual organization Index	Picture-completion test
	Block Design test
	Matrix Reasoning test
3. Working memory Index	Arithmetic test
	Digit span test
	Letter-Numbering sequencing test
4. Processing speed Index	Digit Symbol-Coding test
	Symbol search test

The separate subtest scores of WAIS-III also provides opportunity for *pattern analysis*. In such analysis, the investigator evaluates relatively large differences between subtest scaled scores. For example, a mental disorder like schizophrenia involves poor concentration and impair judgement. As a consequence, such a person would tend to score low on arithmetic test and comprehension test. In fact, Wechsler (1958) had provided a host of pattern analysis tentatively proposed as diagnostically significant.

So far as the psychometric properties of WAIS-III are concerned, it can be used for assessing intelligence of persons aged 16 to 85–89 (Tussky, Zhu & Ledbetter 1997). WAIS-III has impressive reliability and validity when split-half method was used, the average coefficients across age levels were .98 for full scale IQ, .97 for the verbal IQ and .94 for the performance IQ. Test-retest coefficients were only slightly lower, that is, .94, .94 and .88 for FSIQ, VIQ and PIQ respectively. The validity coefficient of WAIS-III in terms of correlation with other tests, particularly with earlier WAIS-R and with the children's version (that is, WISC-III), were high for both individual tests as well as for verbal IQ, performance IQ and full scale IQ. It ranged from .45 to .90.

WAIS-III has some advantages. Its major advantages are as under:

(i) It incorporates the modern multidimensional nature of human intelligence, including fluid intelligence and processing speed.
(ii) It incorporates the possibility of pattern analysis.
(iii) It is appropriate and most suitable for assessing adult human intelligence.
(iv) It uses deviation IQ.
(v) It has impressive degree of reliability and validity.
(vi) It uses a point scale.
(vii) It includes performance scale.
(viii) It makes provision for index score which provides a support to multidimensional nature of human intelligence.

However, WAIS-III also has some disadvantages as under:

(i) It is a poor measure of extreme levels, that is, high or low level of intelligence.
(ii) It does not take into consideration the theories of multiple intelligence as enunciated by Gardner (1983).
(iii) It has poor reliability for the individual subtests.

The latest version of the test is WAIS-IV that was released in 2008. This version is composed of ten core subtests and five supplemental subtests. The scores on the 10 core subtests constitute the Full Scale IQ. In WAIS-IV, the verbal / performance subscales from the previous versions have been deleted and replaced by four index scores as under:

(i) Verbal Comprehension Index (VCI)
(ii) Perceptual Reasoning Index (PRI)
(iii) Working Memory Index (WMI)
(iv) Processing Speed Index (PSI)

In WAIS-IV the General Ability Index (GAI) has been included and it consists of the similarities test, vocabulary test and information test from Verbal Comprehension Index and the Block Design Test, Matrix Reasoning Test and Visual Puzzles Tests from the Perceptual Reasoning Index. The GAI is a good measure of cognitive abilities and is less affected by the impairments of processing and working memory. The WAIS-IV is appropriate for use with persons aged 16–90 years. In WAIS-IV, two broad scores are also generated that are helpful in summarizing the general intellectual abilities. One is Full Scale IQ (FSIQ) based on the total combined performance of VCI, PRI, WMI and PSI and another is General Ability Index (GAI) that is based only on the six subtests that VCI and PRI consists of. The subtests of WAIS-IV are given in Table 9.5.

Table 9.5 Subtests of WAIS-IV

Verbal comprehension	Perceptual reasoning	Working memory	Processing speed
• Similarities (X)	• Block Design (X)	• Digit Span (X)	• Symbol Search (X
• Vocabulary (X)	• Visual Puzzles (X)	• Arithmetic (X)	• Coding (X)
• Information (X)	• Matrix Reasoning (X)	• Letter-Number Sequence (Y)	• Cancellation (Y)
• Comprehension (Y)	• Picture Completion (Y)		
	• Figure Weights (Y)		

X = Core test; Y = Supplemental test

Downward Extension of the WAIS-III: The WISC-III and the WPPSI-R

Wechsler made downward extension of WAIS-III so that it could be applied to the assessment of intelligence of children. For this purpose, Wechsler developed an intelligence scale for children which was first published in 1949, revised in 1974 and most recently revised in 1991. This is known as *Wechsler Intelligence Scale for Children-Third Edition* (WISC-III). It measures intelligence from ages 6 through 16 years, 11 months and 30 days. Like WAIS-III, WISC-III also consists of verbal subtests and performance subtests. In fact, WISC-III contains 13 subtests, three of which are supplementary. The verbal subtests of WISC-III are: Information test, Comprehension test, Arithmetic test, Similarities test and Vocabulary test. Digit span test is a supplementary verbal test. The performance subtests of WISC-III are: Picture completion test, Picture arrangement test, Block design test, Object assembly test and Coding test. Two supplementary performance subtests are Maze test and Symbol search test. All these subtests, with the exception of Maze test and Symbol search test, parallel the corresponding WAIS-III subtests in contents and functions being measured.

In WISC-III, scaled scores are determined on the basis of raw scores through the norms at each age level. Like WAIS-III, here too, the mean scaled score is set at 10 and standard deviation at 3. Scaled scores are summed for obtaining verbal, performance and full scale IQs. These totals are then compared against a single standard score with a mean of 100 and standard deviation of 15 for each of the three IQs.

Reliability and validity coefficients of WISC-III are highly satisfactory. Split-half reliabilities for full score, verbal IQ and performance IQ averaged .96, .95 and .91 respectively. Test-retest reliability coefficients are only slightly below these values. The correlation between the WISC-III and WAIS-III and further downward extension, that is, WPPSI-R were evidences for validity coefficients. All these coefficients for the full scale, verbal and performance IQs are in between .70s and .80s. The WISC-III has also been correlated with the Stanford–Binet scale, with the majority of coefficients in the .60s and .70s for individual subtests and in the .80s to .90s for the three IQs.

As a fair evaluation to WISC-III, it can be said that although it shares many of the strengths of WAIS-III, it has the following problems:

(i) WISC-III fails to incorporate recent developments in cognitive science and theories of multiple intelligence. As we know, cognitive science tends to emphasize analysis in terms of the various components of information processing (Hunt 1980) and abilities like executive processes (Sternberg 1984a), organizing skills (Das 1987) and metacognition (Sternberg & Gardner 1982). We get a meagre reflection of an attempt to incorporate any of these abilities in WISC-III in terms of only as a supplemental symbol search interest. In fact, the reality is that in the 1991 revision of this scale, minimal changes have been introduced.

(ii) WISC-III lacks treatment validity (Witt & Gresham 1985). In fact, specific deficits on WISC-III don't necessarily reflect specific treatment because such deficits are rarely related to remedial intervention for children.

(iii) WISC-III suffers from the problem of selection bias towards certain cultural and racial groups (Saccuzzo & Johnson 1995). Although this scale makes prediction about school achievement of children regardless of ethnic background, the reality is that white children get higher IQ scores than children of other ethnic groups. The obvious result is *selection bias*. African-American children are overselected for the mentally retarded classes whereas they are underselected for the gifted and superior classes.

(iv) WISC-III lacks the index approach of WAIS-III.

The next version of the test is WISC-IV, which was released in 2003 followed by UK version in 2004. This fourth edition of WISC was adapted and standardized for India in 2012. The WISC-IV has 15 subtests, 10 of which formed part of previous WISC III. The five new subtests are: Picture Concepts test, Matrix Reasoning test, Letter-Number Sequencing test, and two supplemental tests, namely, Cancellation test and Word Reasoning test. On WISC-IV a total of five composite scores could be derived: Full Scale IQ (FSIQ) that represented overall intellectual ability, Verbal Comprehension Index (VCI), Perceptual Reasoning Index (PRI), Processing Speed Index (PSI) and Working Memory Index (WMI). The WMI was formerly called as Freedom from Distractibility Index and PRI was formerly known as Perceptual Organization Index. The subtests included by the four indices are as under:

Verbal Comprehension Index

- Vocabulary (X)
- Similarities (X)
- Comprehension (X)
- Information (Y)
- Word Reasoning (Y)

Perceptual Reasoning Index

- Block Design (X)
- Picture Concepts (X)
- Matrix Reasoning (X)
- Picture Completion (Y)

Working Memory Index

- Digit Span (X)
- Letter-Numbering Sequencing (X)
- Arithmetic

Processing Speed Index

- Coding (X)
- Symbol Search (X)
- Cancellation (Y)

'X' indicates core subtests whereas 'Y' indicates supplemental tests.

Towards the fall of 2014, the latest version of WISC is the WISC-V.

In 1967, Wechsler developed another scale for children between 4 and 6 years of age. This was named as *Wechsler Preschool and Primary Scale of Intelligence* (WPPSI). Thus WPPSI is a further downward extension of WISC. WPPSI was revised in 1989 and called WPPSI-R and covers the range of 3 years to 7 years and 3 months. Like WAIS-III and WISC-III, it has verbal subtests and performance subtests. In fact, WPPSI-R parallels the WAIS-III and WISC-III in the format, in the method of determining reliabilities and subtests, too. However, only two different subtests are included in WPPSI-R: *Animal Pegs test* and *Sentences test*. In the Pegs test, children are required to place a coloured cylinder into an appropriate hole in front of an animal within a specified time, whereas in the sentence test, they are asked to repeat sentences told by the examiner. Both these tests are optional tests.

WPPSI-R has also a sufficient degree of reliability and validity. Reliability coefficients are comparable to those obtained with other Wechsler scales. Validity coefficients in form of correlation against the measure of Stanford–Binet scale was .92.

The third edition of WPPSI called WPPSI-III was released in 2002. Its current edition, that is, fourth edition called WPPSI-IV was published in 2012 and is considered as an innovative measure of cognitive development for preschool and young children coming in between the age range of 2 years 6 months to 7 years 7 months. WPPSI-IV includes some content changes like new processing speed measures, the addition of working memory subtests and an expanded factor structure that includes separate visual spatial and fluid reasoning composites. The test structure of WPPSI-IV includes three level of interpretation: Full Scale level, Primary Index Scale level and Ancillary Index Scale level. The details are as under:

Primary Index Scale

(i) Verbal Comprehension Index (VCI)
(ii) Visual Spatial Index (VSI)
(iii) Working Memory Index (WMI)
(iv) Fluid Reasoning Index (FRI)
(v) Processing Speed Index (PSI)

Ancillary Index Scale

(i) Vocabulary Acquisition Index (VAI)
(ii) Nonverbal Index (NVI)
(iii) General Ability Index (GAI)
(iv) Cognitive Proficiency Index (CPI)

Clinical psychologists, school psychologists and neuropsychologists working in schools, hospitals, clinics, universities use WPPSI-IV for evaluating children for cognitive delays, intellectual disabilities, autism and giftedness, etc.

Concluding remarks on Wechsler Scales

All the three Wechsler scales, namely, WAIS-IV, WISC-IV and WPPSI-IV have definitely improved as compared to the decades when they had been originally developed. As a consequence, they have become very popular. Many computer-assisted interpretative programmes and interpretive guides have also added glory to their popularity (Kaufman 1994; Nicholson & Alcorn 1994. However, some critics have also expressed their apprehension that even the improved versions of the Wechsler scales would soon become obsolete in light of the current demands for links between assessment instruments / tests and intervention strategies (Shaw, Swerdlik and Laurent 1993; Sternberg 1993). In view of this, it can be said that the weakest point of all the three Wechsler scales has been their lack of theoretical grounding which makes it hard to provide a coherent basis of interpretation.

A comparison between Binet Scales and Wechsler Scales

There are two profound differences between Binet scales and Wechsler scales. These differences can be discussed as under:

(1) Binet scales are age scales whereas Wechsler scales are point scales. Binet scales (from 1908 to 1972) are basically age scales because in this scale, items are grouped by age level and in each age level two thirds to three fourths of individuals are expected to pass a group of tasks included by each age level. In age scale, arrangement of items is no way concerned to their content. At any age level, there may be different tasks related to different aspects of the content. For example, at a particular age level one task may be related to perceptual speed, another task to memory and still another task to

concentration or language skills. At another age level, one task may be related to reasoning. In this way, different types of content may be scattered throughout the scale. Besides, in age scale the subjects don't receive a specific amount of credit or points for each task completed. For example, if Binet scale required that the subject must pass four out of five items in order to receive a credit, and suppose he passes only three items, he will receive no credit for this part of the test.

Wechsler scales are point scales where credits or points are assigned to each item. Here the subject receives the specific amount of credits or points for each item passed. The Binet scale has never given such credit. On the modern Binet scale, a basal score may be arrived, whether the person is correct on three out of four or on all four items at two consecutive levels. The person who gets three items correct obtains the same basal score as the one who gets all four correct. In this way it can be said that although modern Binet scale has abandoned the concept of age scale, it has not yet adopted the point scale concept.

(2) Binet scale and Wechsler scale differ from each other on the basis of performance scale concept. In the Binet scale, there is emphasis upon verbal skills and language for which it has been criticized from time to time. Wechsler was aware of this problem and so he included performance scale in this test. In the performance scale, the subject is required to do something and that becomes the basis for his intelligence.

Although early Binet scales contained some performance tasks, these were mostly restricted to younger age levels. Apart from this, results of the subject's response to a performance task on the Binet scale were extremely difficult to separate from the results for verbal tasks. As a consequence, it was very difficult to tell to what extent a subject's response to a performance task increased or decreased the total score on the test. In Wechsler scale, the picture is entirely different because it contains two separate scales—one for verbal task, called verbal scale, and another for performance task, called performance scale. In fact, Wechsler intelligence scale was the first to offer the possibility of directly comparing an individual's verbal and nonverbal intelligence. That is, verbal scales and performance scales were standardized on the same sample and the results of both scales were manifested in comparable units.

In this way, we find that the Binet scales and Wechsler scales differ from each other on two major counts.

The Kaufman Scales

Developed in the 1980s and 1990s, the Kaufman scales tend to incorporate recent developments in test construction. The Kaufman scales are individually administered tests, designed for many different uses. There are three types of intelligence tests included in Kaufman's scales:

(1) Kaufman Assessment Battery for Children (K-ABC)

(2) Kaufman Adolescent and Adult Intelligence Scale (KAIT)

(3) Kaufman Brief Intelligence Test (K-BIT)

These scales may be discussed as under:

(1) *Kaufman Assessment Battery for Children* (K-ABC): The K-ABC is an individual intelligence test for children between $2\frac{1}{2}$ and $12\frac{1}{2}$ years of age (Kaufman & Kaufman 1983a, 1983b). The test consists of 16 subtests combined into five global scales called *sequential processing, simultaneous processing, mental processing composite* (a combination of sequential and simultaneous processing), *achievement* and *nonverbal*.

In fact, the K-ABC measures intelligence through its mental processing scales, namely, sequential processing, simultaneous processing and mental processing. Theoretically, the K-ABC is based upon several types of approaches (Kaufman 1984). For example, as its base, it includes the neuropsychological model of brain functioning by the renowned Russian neuroscientist Luria (1966), the theory of split brain functioning of Sperry (1968) and the theories of information processing of the famous cognitive psychologist, Neisser (1967). In the works of these scientists, Kaufman (1984) clearly noted distribution between two types of higher-brain processes which was referred to as sequential-simultaneous distinction. Sequential processing refers to the children's ability to solve problems by mentally analyzing inputs into a serial order or sequential order as we find in case of number and word-order recall. Simultaneous processing, which takes place in parallel, refers to children's ability to synthesize and integrate information for solving a problem (A S Kaufman, R W Kaufman, and M L Kaufman, 1985). In K-ABC, the major purpose of providing separate measures of simultaneous and sequential processing is to identify the child's unique strength and problem-solving strategies that can help others develop remedial intervention strategies for a child.

Besides sequential-simultaneous distinction, the K-ABC also provides for an independent achievement score. In fact, offering an independent measure of intelligence (through sequential-simultaneous processing) and achievement is one of the major advantages of K-ABC. In addition, K-ABC also includes a nonverbal scale which provides a measure of ability of children who are linguistically different and handicapped (Flanagan 1995; Reynolds & Kamphaus 1997).

The obtained raw scores on each of 16 subtests are converted to standard scores with a mean of 10 and standard deviation of 3. The global scale score can be converted into a standard score with a mean of 100 and standard deviation of 15, percentile and age norms.

Reliabilities and validities are also sufficient. Split-half reliabilities range from .72 to .88 for preschool children and from .71 to .85 for school-age children. Likewise, split-half coefficients for the global scales range from .86 to .93 for preschool children and .89 to .97 for school-age children. Test-retest coefficients are slightly lower than these coefficients. Test manual reports a variety of validity data which have received considerable support. For example, factor-analytic studies support its sequential-simultaneous and mental-processing-achievement distinctions (McEsters et al. 1998).

The K-ABC has been criticised. Jensen (1984) has pointed out that this test as compared to Binet and Wechsler scales has poorer predictive validity for school achievement as well as less effective assessment of general intelligence. Other criticisms relate to the K-ABC's imperfect model with its theoretical foundation and disproportionate contribution of mental processing composites to intelligence. Besides, this test is highly visual which limits its use with visually handicapped children and it also does not measure the extremes of intelligence (Alken 1987). Sternberg (1984b) has also criticised this test bitterly. He has pointed out that the test suffers from a noncorrespondence between its definition and its measurement of intelligence. He has also questioned the empirical support for the theory underlying K-ABC. He has further pointed out that the test gives overemphasis on rote learning at the expense of the ability to learn.

The second edition of KABC called KABC-II was published in 2004 and it measures the processing and cognitive abilities of children as well as adolescents in between ages 3 to 18. KABC-II differs in its conceptual framework and test structure from original KABC. It is said that while KABC is tied to simultaneous/sequential processing approach, the KABC-II covers two different theoretical models: the Cattell–Horn–Carroll (CHC) psychometric model of abilities and Luria's neuropsychological theory of processing.

KABC-II has 18 subtests of two types: core and supplemental. Before the testing is begun, the examiner has to decide which model to follow: CHC or Luria. Luria's model consists of four subscales: simultaneous processing scale, sequential processing scale, learning ability and planning ability. CHC model renames simultaneous processing scale as visual processing (GV) sequential processing scale as short-term memory (Gsm), learning ability as long-term storage and retrieval (Glr) and planning ability as fluid reasoning (Gf). In CHC model, an additional, fifth scale called crystallized ability (Gc) or knowledge has also been included.

KABC-II includes the following subtests:

1. Simultaneous / Gv

- Triangles
- Face reognition
- Block counting
- Conceptual thinking
- Rover
- Gestalt closure
- Pattern reasoning (ages 5 and 6)
- Story completion (ages 5 and 6)

2. Sequential / Gsm

- Word order
- Number recall
- Hand movements

3. Planning / Gf

- Pattern reasoning (ages 7–18)
- Story completion (ages 7–18)

4. Learning / Glr

- Atlantis
- Atlantis delayed
- Rebus
- Rebus delayed

5. Knowledge or Crystallized ability (Gc) (only in CHC mode)

- Riddles
- Expressive vocabulary
- Verbal knowledge

KABC-II yields two general intelligence composite scores: Mental Processing Index (MPI) based on Luria's model and Fluid-Crystallized Index (FCI) based on CHC model. In general, Luria model takes about 25–60 minutes in its administration whereas the CHC model takes about 30–75 minutes, depending upon the age of the child. This test helps the researcher in identifying a person's strengths and weaknesses in cognitive ability and mental processing.

(2) *Kaufman Adolescent and Adult Intelligence Test* (KAIT): The KAIT was constructed as a measure of intelligence for persons whose ages ranged from 11 to 85 years or older A S Kaufman & N L Kaufman 1993). In fact, this test represents an attempt to integrate the theory of fluid and crystallized intelligence formulated by Horn & Cattell (1966) with notions about adult intelligence proposed by other theorists (Luria 1980; Piaget 1972). This test is composed of a crystallized scale and a fluid scale. The former measures concepts acquired from schooling and acculturation whereas the latter measures the ability to solve new problems.

In fact, KAIT incorporates two components: a core battery and an expanded battery. The core battery comprises fluid intelligence scale, crystallized intelligence scale and six subtests. It takes about 1 hour and 15 minutes to complete. The expanded battery includes core battery tests as well as four additional subtests and takes about 1 hour and 30 minutes in its completion. The core battery subtests related to fluid intelligence are: logical steps, a test of sequential reasoning, a test of measuring induction, mystery codes, Rebus learning and a test of long-term memory. The core battery subtests related to crystallized intelligence are: a test of word knowledge and language development, definitions, double meanings, a measure of language comprehension, auditory comprehension and a test of listening ability.

The expanded battery also incorporates memory for block designs, a measure of visual processing related to fluid intelligence, famous faces, a test of cultural knowledge related to crystallized intelligence, Rebus delayed recall as well as auditory delayed recall.

In KAIT battery, there is an optional supplemental mental status examination. Since KAIT yields IQ scores in a relatively wider range, that is, from much lower-than-average intelligence to much higher-than-average intelligence, it is often used as an assessment of individuals with exceptional abilities like gifted children.

(3) *The Kaufman Brief Intelligence Test (K-BIT):* K-BIT was designed as a quick screening test for estimating the level of intellectual functioning (Kaufman & Kaufman 1990). It covers the age range of 4 to 10 years. It is an individually administered test but is not a shortened version of either the K-ABC or the KAIT. It consists of verbal scale, nonverbal scale and a composite. The verbal scale consists of one verbal subtest of 45 expressive vocabulary items and 37 definitions and one nonverbal scale of 48 matrices. The three scores, namely, verbal score, nonverbal score and the composite score are expressed in deviation IQ units. It has a higher reliability coefficient.

The second edition of KBIT called KBIT-2 was published in 2004. This is a brief intelligence test for individuals from ages 4 to 90 years. KBIT-2 consists of verbal subscales, nonverbal subscales and IQ composite. The verbal scale comprises two combined subtests, which assess receptive vocabulary and general information as well as reasoning, comprehension and vocabulary knowledge (Riddles). The nonverbal scales incorporate Matrices subtest to tap the ability to complete visual analogies and understand relationship. KBIT-2 requires 15 to 30 minutes for its completion. Adults, gifted children, as well as examinees with reflective style of test taking may take a longer time. It is also worth noting that none of three subtests of KBIT-2 are timed.

Nonverbal Intelligence Scales

In their efforts to construct intelligence tests which can be applied across cultures, psychometricians have developed many nonverbal scales which can be administered either individually or in a group. Two important such scales to be discussed here are: *Raven's Progressive Matrices* (RPM), and *Goodenough–Harris Drawing test.*

The Raven's Progressive Matrices (RPM) test is one of the best known and most popular intelligence tests that can be used either in a group or individually. It covers people from 5 years to elderly adults. It consists of 60 multiple-choice questions listed in order of difficulty (Kaplan & Saccuzzo 2009). It has been designed primarily as a measure of Spearman's *g* factor of general intelligence (Raven 1983; J Raven, J C Raven & Court 1995). The items of the test consists of a set of matrices or arrangements of designs into rows and columns, from each of which a part remains missing. The task of the subject is to choose the missing insert from the given alternatives for completing a pattern. Many patterns are presented in form of 6 ×6, 4 ×4, 3 ×3 or 2 ×2 matrix that gives the test its name. The easier items simply require accuracy of discrimination but the difficult items require some complex processes like analogies, permutations, alterations of patterns and other logical relations. The test is usually administered with no time limits.

The RPM is available in three different types of forms: *The Standard Progressive Matrices* (SPM-1996 Edition), the *Coloured Progressive Matrices* (CPM-1990 Edition), and the *Advanced Progressive Matrices* (APM-1994 edition). SPM is suited for average individuals between the ages of 6 and 80 years. CPM is available for younger children or special persons, especially mentally retarded ones who cannot be tested with the SPM for one or the other reasons. APM has been developed for above-average adolescents and adults.

Reliabilities and validities of RPM were high and satisfactory. Test-retest reliabilities ranged from .70 to .90 and internal consistency coefficients are largely in the .80s and .90s. Validity coefficients with both verbal and performance tests of intelligence range between .40 and .75, providing a higher value with performance tests than with verbal tests.

Goodenough Draw-a-Man test is another nonverbal measure of intelligence. This is a very simple test in which the examinee is simply instructed to make a picture of a man as accurately as he can. This test was originally developed in 1926 and was used without change until 1963. In this year, it was revised by Harris and now, the title became Goodenough–Harris Drawing Test. In this revised form, like in the original test, the emphasis in drawing is placed upon the child's accuracy of observation and on development of conceptual thinking rather than on any artistic skills. Subjects get credit for each item included in their drawings. As a rule each detail is given one point, with possibility of a total of 70 points. In this revised form, there are three scales: *The Man scale, The Woman Scale and the Self scale.* In the Man Scale, the examinee is asked to draw a picture of a man; in the Woman Scale, the examinee is asked to draw a picture of a woman. In the Self Scale, which is basically treated as a projective test of personality, the examinee draws a picture of himself.The point scores on each of the three scales are translated into standard scores with a mean of 100 and standard deviation of 15. The test is administered to ages 3 years to 15 years 11 months. However, the preferred age is 3 to 10 years.

This test has sufficient degree of reliabilities and validities. The test-retest reliabilities, as well as split-half and scorer reliability, are adequate (Dunn 1967; Harris 1963). The validity of the test is provided by correlations with other intelligence tests and although these correlations vary widely, the majority are above .50.

All these nonverbal scales and similar others, although initially developed for cross-cultural testing, have also been widely and popularily used in clinical and counseling settings. As already said, these nonverbal tests can be used either in individual situations or in groups. Besides these, there are some exclusive group tests which will be briefly discussed here.

History bears testimony to the fact that mass testing began during World War I with development of the *Army Alpha test* and the *Army Beta test* for use in the United States Army. The Army Alpha test was a verbal test designed for the purpose of general screening and placement purposes, whereas the Army Beta test was a nonlanguage test for use with persons who could not be properly tested due to illiteracy or foreign-language background. The pattern established by these tests was closely followed in the subsequent development of many group tests for civilian applications.

Three group tests of intelligence are worth mentioning here: *Kuhlmann–Anderson test, Henson–Nelson test* and *Cognitive abilities test.*

The Kuhlmann–Anderson test (KAT) is a group intelligence test which measures intelligence covering kindergarten through 12th grade. In fact, it covers eight separate levels in between kindergarten and 12th grades and it each level contains several tests with a variety of items on each level. KAT items are nonverbal and it is suited not only to young children but also to those who might be handicapped in verbal expressions. The results of the latest edition of the KAT can be expressed in verbal, quantitative and total scores. At some levels, total scores can be expressed as deviation IQs. Scores at the other levels can be expressed in the form of *percentile band* which is like a confidence interval providing the range of percentiles that most likely represent a subject's true score.

Reliability coefficients, particularly split-half coefficients, are in low .90s and test-retest coefficients range from low .80s to the high .90s. So far as the validity is concerned, the KAT correlates highly with different intelligence tests. In sum, it can be said that the KAT is an extremely sound group nonverbal test of intelligence.

Henmon–Nelson test is another group mental ability test for all grade levels. This test produces a single score which is believed to reflect general intelligence. Two sets of norms are available. One set is based on raw score distributions by age and the other set on raw score distributions by grade. There are provisions for raw scores that can be converted into deviation IQs as well as percentiles. It consists of 90 items which takes about 30 minutes in its completion. The reliability coefficients, particularly split-half and test-retest, run in .90s. As a validity measure, the test correlates well with a variety of intelligence tests and these correlation coefficients range from .50 to .84. The major advantage of this test is that it can readily help predict future academic success. However, by providing only a single score related to Spearman's *g* factor, this test does not consider multiple intelligence. Likewise, when the test was being developed, no special effort had been made to check for content bias, either by judging or by statistical analysis. Another problem with this test is its relatively *low ceiling*. For example, to achieve an IQ of 130 a ninth-grade child would have to answer about 85 out of 90 items correctly. This automatically leaves only five items to discriminate all those above 130.

The Cognitive Abilities Test is another important group intelligence test which provides three separate intelligence scores: *quantitative, verbal* and *nonverbal*. This test is designed for measuring intelligence of poorly educated people as well as for people for whom English is a second language. The test authors have taken special care for eliminating cultural bias and for eliminating the effect of test-taking skills, and its test administration includes extensive practice exercises. Each of the three subtests of cognitive abilities test requires 32 to 34 minutes of actual working time which should be spread over two to three days. The reliabilities, that is K-R$_{20}$, for the verbal test are in high .90s, for the quantitative test in low .90s and for the nonverbal in high .90s.

On the negative side, it can be said that the three tests of cognitive abilities test are primarily power tests but no data are provided in its support. Likewise, uncertainty remains as to whether norms which apparently appear strong, are representative (Kaplan & Saccuzzo 2001).

Some Indian Intelligence Tests

Many types of intelligence tests have been developed in India, a glimpse of which can be had by examining a selected number of abstracts. One of the pioneers of intelligence testing in India is Professor S M Mohsin who constructed an intelligence test in Hindustani in the late 1930s (Pandey 2000). In The first *Mental measurement Handbook for India*, a list of 103 tests of intelligence in different Indian languages including the test of general intelligence by Professor Mohsin has been provided (Long & Mehta 1966). Samples, reliability, validity and norms were highly satisfactory. In fact, documentation of Indian tests is done by *National Library of Educational and Psychological tests* (NLEPT) at the National Council of Educational Research and Training (NCERT). NLEPT has published an Indian mental measurement handbook on intelligence tests. Some of the popular Indian intelligence tests are General Mental Ability test for children by R P Shrivastava and K Saxena, Group test of intelligence by P Ahuja, Group test of Intelligence by G C Ahuja, Verbal Intelligence Scale by R K Ojha and Ray Chaudhary, Test of General Intelligence for college students by S K Pal and K S Mishra, Group Test of Intelligence by R K Tandon, Emotional Intelligence Scale by A Hude, S Pethe and Upindar Dhar and Social Intelligence Scale by N K Chadha and Usha Ganesan, Indian child Intelligence test by Usha Khire. These inteligence tests are marketed by the National Psychological Corporation, Agra.

Kumar (1991) has provided a decadewise analysis of development and nature of intelligence tests in India. He reported that approximately 70 per cent intelligence tests were constructed and standardized during the 1960s and 1970s. However, there occurred a decline in the 1980s. Of these tests, 67 per cent were verbal, 23 per cent nonverbal and 10 per cent were performance-based.

51 per cent of these tests were developed in Hindi and they were group tests meant for assessing intelligence of 13–17 years of age.

TYPES OF INTELLIGENCE TEST SCORES

Usually, there are three types of intelligence test scores which are very common: the mental age, the intelligence quotient and the deviation intelligence quotient. A detailed discussion of each of these intelligence scores is as under:

Mental Age

The concept of mental age is very popular in the measurement of intelligence and it was for the first time introduced by Binet, a French psychologist, in 1908, when he published the first revision of his intelligence test—the Binet–Simon scale. Mental age, in the words of Tuckman (1975, 329) refers to "a score that is determined by comparing a child's score with the scores obtained by younger and older children in the norming group." Thus mental age is a score which is obtained by comparing a child's obtained score with the average score (called age norms) at his age level. For example, suppose a 10-year-old child obtains a score of 60 on a certain intelligence test, which is an age scale. Further, suppose that the mental age norms on this test indicate that a child in order to have the mental age of 9 and 10 years must obtain a score of 60 and 70 respectively. In such a situation his mental age will be 9 years (and not 10 years), although his real age is 10 years. Likewise, there may be another child who is 8 years old, but obtains a score of 60 on that test. In such a situation, he will be said to have a mental age of 10 years although his chronological age is only 8 years. Mental age is usually expressed in terms of years and months such as 6 years 2 months, 6 years 6 months, 8 years 6 months, and so on. The most common way to derive mental age is to obtain a representative sample at each of the age levels for which the intelligence test is being developed. For each age level, items are prepared and given to them. The mean (or average) performance of each age level becomes the mental age for that age-mate.

Intelligence Quotient

As mentioned earlier, the concept of intelligence quotient (or IQ) was for the first time introduced in the 1916 American revision of the Binet–Simon scale done at Stanford University by Terman. IQ is defined as the ratio of mental age to the chronological age or real age, multiplied by 100 to avoid decimal fractions. It is, therefore, also known as the ratio IQ. The equation for IQ is:

$$\mathrm{IQ} = \frac{\text{Mental age}}{\text{Chronological age}} \times 100 \qquad (9.1)$$

Suppose a 10-year-old child has mental age of 120 months, then his $\mathrm{IQ} = \frac{120}{120} \times 100 = 100$ which would be considered as normal or average in the sense that his mental age and chronological age are exactly equal. However, if another 10-year-old child has a mental age of 156 months, his $\mathrm{IQ} = \frac{156}{120} \times 100 = 130$ which obviously exceeds the average (that is, 100). Such a child would be considered to be of very superior intelligence in the sense that his performance is much above the performance of his age-mates and is comparable to children three years older than himself. If, however, a 10-year-old child has a mental age of 96 months only, his $\mathrm{IQ} = \frac{96}{120} \times 100 = 80$, which obviously falls below the average, and he will be considered as dull in the sense that his performance is below the average of his age-mates and is comparable to children two years younger than him. Many psychologists have worked over the concept of IQ in different situations. As a result, a consensus (setting aside some minor variations) has emerged for classifying persons on the basis of IQs as provided in Table 9.6.

Although the concept of the IQ ratio has been very popular in measuring intelligence, it has some important shortcomings.

1. It is a known fact that after 18 or 19, the mental age of a person tends to stabilize with a few rare exceptions. Later, the chronological age continues to increase and the mental age remains more or less static. In other words, the numerator of the IQ equation remains constant while the denominator continues to inflate. The natural outcome of this will be that a person *seems* to be less and less intelligent as he or she grows older in age.

Table 9.6 A General classification of person in terms of IQ

IQ	Classification
Above 140	Genius
130–140	Very superior
120–129	Superior
110–119	Bright
90–109	Average
80–89	Dull
70–79	Borderline
69 and below	Mentally defective

2. Another shortcoming, as has been pointed out by many psychologists, is that the variability in IQ score from one test to another is not the same. Naturally, then, on a test with less variability an individual will have to obtain a higher score for being as far above average or normal as on a test with greater variability. In such a case, the IQ score of the different tests would not be directly comparable.

3. It has been commonly seen that the variability in IQ scores for the different age levels on the same test is not the same. In other words, the standard deviations of the IQ scores are not the same for the different age levels on the same test. In such a case, the IQ score would be a misleading index because this would indicate that a person's IQ score varies as he grows older though his position compared to his age-mates remains the same. Thus a 10-year-old child's IQ score of 120 (where SD is 12) may mean a different thing when it is compared with a 7-year-old child's IQ score of 120 (where SD is 9). Thus, the meaning of the same IQ score of 120 at the two age levels with different standard deviations would not be identical (whereas it should be).

The IQ ratio was a very popular concept until about 1960. After that it was practically replaced by the concept of deviation IQ or DIQ so that the above shortcoming, particularly the third one, could be removed. Nowadays the IQ ratio score is easily converted into a normalized standard score with a fixed mean of 100 and fixed SD of 15 or 16. Such a converted IQ ratio score is known as the deviation IQ score or DIQ score. This fixed mean and SD remain constant through all the age levels (for which the test is meant) and for all levels of a single test. If a child's score is one SD above the mean his DIQ score would be 100 + 15 or 16 = 115 or 116, which would indicate that his DIQ lies above the average of his age-mates. Likewise, if his score is one SD below the mean, his DIQ score would be 84 or 85, which would indicate that his DIQ falls below the average of his age-mates. Thus in DIQ the standard deviation has a constant value at each age level and thereby, controls the aberrant variability found in the IQ ratio. A discussion of the DIQ score has already been given in Chapter 7.

DISTINCTION BETWEEN APTITUDE TEST AND ACHIEVEMENT TEST

The terms 'intelligence', 'aptitude' and 'achievement' are commonly used in the field of psychological and educational testing. It is, therefore, essential to make a distinction among these terms as well as among 'intelligence test', 'aptitude test', and 'achievement test'. It is very difficult, if not impossible, to distinguish between the two terms intelligence and aptitude. As discussed previously, intelligence refers to a general set of mental abilities. On the basis of these mental abilities, a child's performance in the future can be predicted. Aptitude, on the other hand, is a person's ability, acquired or innate, to learn or develop knowledge of a skill in some specific area. Freeman (1962, 431) has defined aptitude as "a combination of characteristics indicative of an individual's capacity to acquire (with training) some specific knowledge, skill or set of organised responses such as the ability to speak a language, to become a musician, to do mechanical work." Likewise, Tuckman (1975, 475) has defined aptitude as "a combination of abilities and other characteristics, whether native or acquired, known or believed to be indicative of an individual's ability to learn or to develop proficiency in some particular area." Obviously, both Freeman and Tuckman intend to say more or less the same thing regarding the meaning of aptitude, which, according to them, refers to the capacity or ability to acquire skill or knowledge in a particular area. On the basis of such abilities, the future performance of a child can be predicted. Achievement refers to what a person has acquired or achieved *after* the specific training or instruction has been imparted. In other words, achievement tests are primarily designed to measure the effects of a specific programme of instruction or training (Anastasi 1968). Thus, the performance on the achievement test indicates the performance under known and controlled conditions (because the performance is the outcome of specific training given in a specific field), whereas aptitude tests indicate performance that is the product of multiplicity of experiences of everyday life and therefore, indicate the effect of learning under unknown and relatively uncontrolled conditions.

When the meanings of intelligence, aptitude and achievement are known, it is easier to distinguish among intelligence test, aptitude test and achievement test. Freeman (1962, 431) has defined aptitude test as one "designed to measure a person's potential ability in an activity of a specialised kind and within a restricted range." Achievement test, also known as proficiency test, is one which measures the extent to which a person has acquired or achieved certain information or proficiency as a function of instruction or training (Tuckman 1975). In attempting to distinguish among these tests, intelligence tests and aptitude tests may be placed on one side and achievement tests on the other. Aptitude tests (as well as intelligence tests) are future oriented whereas achievement tests are present and past oriented (Stanley & Hopkins 1978, 361). The primary purpose of an aptitude test (and intelligence test) is to predict what a person *can* learn whereas the primary purpose of an achievement test or proficiency test is to evaluate what a person *has learned*. In other words, intelligence test and aptitude test are *psychometric* tests whereas achievement tests are *edumetric* tests. A psychometric test is one whose major intention is to measure individual differences like a general ability, aptitude or trait, etc., and an edumetric test is one whose major intention is to measure the gain or growth of individuals such as measurement of skill, proficiency and achievement (Carver 1974, 513). However, a single test can be used either in a psychometric way or edumetric way depending upon the purpose of the test. For example, when a numerical aptitude test is used for the purpose of predicting the future numerical ability of a person, it would be considered to be a psychometric test but when the same test is used to measure the numerical ability of a person as a function of some training, it would be considered to be an edumetric test. Psychologists, in general, use the test in a psychometric sense and educators, in general, use the test in an edumetric sense.

The above distinction between aptitude test and achievement test is not very clear-cut in actual practice. In a recent study, Bracht and Hopkins (1972) have demonstrated that an educational achievement test (obviously, a kind of achievement test) predicts future achievement better than the intelligence test. Now, psychometricians are using a new term, *"test of developed ability"* in place of aptitude tests and achievement tests. The assumption underlying the use of the term is that all general intelligence tests, aptitude tests (multiple aptitude batteries and special aptitude tests) and achievement tests are ability tests and simply measure the developed ability of a person in one or more field.

TYPES OF APTITUDE AND ACHIEVEMENT TESTS

Aptitude Tests

Aptitude tests may conveniently be grouped into two categories—multiple aptitude tests and special aptitude tests. Multiple aptitude tests are those which intend to measure several aptitudes, each by an independent subtest. Hence, multiple aptitude tests are not tests but rather *batteries of* tests. Special aptitude tests are those which intend to measure only one aptitude. They appeared earlier than multiple aptitude tests. Among the earliest special aptitude tests was the mechanical aptitude test. Special aptitude tests were developed at the time when the primary emphasis in testing was placed on the general intelligence test and it was thought that for describing the individual measurement of intelligence, it must be supplemented by the measurement of special aptitudes like mechanical, numerical, etc. But later researches, particularly the factor-analytic researches, conducted by Thurstone and Guilford revealed that intelligence itself consists of several independent special aptitudes and since then, the emphasis has shifted from the special aptitude test to the multiple aptitude test. Not only this, several aptitude tests are now easily included in multiple aptitude batteries. In the following sections, some of the aptitude tests representative of the above two broad categories are discussed.

Differential Aptitude Tests

The Differential Aptitude Test or DAT is one of the most common multiple aptitude tests. First published in 1947, the battery has undergone several revisions and restandardizations and presently, it is available in its fifth edition done in 1992. The battery has been developed by Bennett, Seashore and Wesman and comprises eight subtests—*verbal reasoning, numerical reasoning, abstract reasoning, mechanical reasoning, clerical and speed and accuracy, space relations, spelling, and language usage*. The battery is mainly meant for educational and vocational counselling of students from grade 8 through 12. The whole battery has two equivalent forms, S and T, and it roughly takes about three hours to administer. Scores on each subtest are converted into the percentile ranks for their proper interpretation. In addition to eight subscores, a ninth subscore is provided by adding the scores on the verbal reasoning (VR) test and numerical reasoning (NR) test and, subsequently, the composite score is transformed into the percentile rank. The VR + NR score becomes the index of general scholastic aptitude, which is interpreted as one index of mental ability. The DAT has also been adapted by Indian psychologists to suit local requirements.

General Aptitude Test Battery (GATB)

The General Aptitude Test Battery (GATB) was developed by the United States Employment Service in 1962 for use primarily in the armed services. The battery was constructed as a function of extensive factor analysis in which 59 tests were correlated. On the basis of the factor analysis, 10 factors were chosen which, in a subsequent revision of the GATB, were reduced to nine to be measured by 12 tests. The nine factors of the GATB are: *intelligence* (G), *numerical aptitude (N), verbal aptitude (V), spatial aptitude (S), form perception* (P), *clerical perception* (Q) *motor co-ordination* (K), *finger dexterity* (F), and *manual dexterity* (M). Of the 12 tests developed for measuring these nine factors, eight are verbal tests and four tests developed for measuring *F* and *M* are nonverbal tests requiring simple apparatus for subjects. The tests intended to measure factor G through Q, are available in alternate forms. The scores obtained on the nine tests of the GATB can be converted into a standard score with a mean of 100 and standard deviation of 20 for their interpretation. The whole battery takes about two hours and thirty seconds in its complete administration. The GATB has been widely used in the employment services, too. The primary limitation of the battery is that all its tests are highly speeded and some important aptitudes have not been included. For example, test of mechanical reasoning has not been included in the battery.

Flanagan Aptitude Classification Tests

The Flanagan Aptitude Classification Test or FACT is another battery of the multiple aptitude tests. The battery has been primarily developed for vocational counselling and employee selection and has been named after its author, Flanagan, who did the job analysis of several occupations, which led to the emergence of 21 'job elements' or abilties that discriminated well between successful and unsuccessful workers on each job. For measuring 19 job elements out of 21, verbal tests have been developed and for the remaining two, performance tests have been developed. The whole battery requires three testing sessions and more than ten and a half hours are required in its complete administration.

Armed Services Vocational Aptitude Battery (ASVAB)

The Armed Services Vocational Aptitude Battery (ASVAB) is another popular multiple aptitude battery. It is designed for students in grades 11 and 12 as well as for postsecondary schools. It provides scores used in both educational and military settings. When used in military settings, ASVAB results can help identify such students who are fit for various types of military occupational programmes.

ASVAB comprises ten subtests: *general science, paragraph comprehension, arithmetic reasoning, work knowledge, numeral operations, coding speed, auto and shop information, mathematics knowledge, mechanical comprehension* and *electronics information*. These ten subtests are grouped into various composites. There are three types of academic composites: academic ability, verbal ability and mathematical ability. Likewise, there are four occupational composites—mechanical and crafts, business and clerical, electronics and electrical, and health and social. Besides these, there is also an overall composite reflecting general ability. The overall psychometic characteristics, that is, its reliability, validity, norms, etc., are excellent (Ree & Caretta 1994, 1995).

Besides the above multiple aptitude batteries, there are some special aptitude tests. Although nowadays not much emphasis is being given to the development of special aptitude tests, for some reason they are still being preferred. *First*, special aptitude tests provide more flexibility in the choice of the appropriate and relevant test than the multiple aptitude batteries. *Second*, there are certain areas of aptitudes, which are seldom included and well covered by the multiple aptitude batteries. Artistic talents, musical talents, creative talents, vision, hearing and other areas requiring some motor functions are examples of the common areas rarely included and covered by the aptitude batteries. Some of the important *special aptitude tests* are presented below.

Sensory Tests

Among the early psychological tests, sensory tests devised by Galton and others to measure intelligence were first in vogue. But these tests could not gain prominence due to their failure to predict intellectual accomplishment. Researches on the sensory capacities, however, were continued since that time and today, there are sophisticated techniqes for measuring visual and auditory sensitivity. Vision is not a unitary capacity. It has several dimensions and a person who is normal in one dimension, may be impaired in another. For example, a person may be colour blind but his vision may be perfectly normal otherwise. Among the visual dimensions which have attracted the attention of psychologists for extensive studies are the visual near acuity activity (usually measured at a reading distance of 15 to 18 inches) and far acuity (usually measured at the distance of 20 feet), the visual distance or depth perception, the colour discrimination and the phobia (that is, muscular control of eyes). The most common method of measuring far acuity is the wall chart. One such chart is known as the *Snellen chart*, which consists of rows of letters of gradually decreasing size. The Snellen chart is usually placed at a distance of 20 feet and the person is asked to read the letters. The three most important and well-known instruments for measuring each of the above remaining dimensions are the Ortho-Rater, the American Optical Sight Screener, and the Keystone Telebinocular. These three instruments are the most common measures in the selection of personnel in industry and are also used for screening purposes in schools.

Audition is another important sensory field where psychologists have concentrated their attention. Like vision, audition is also considered as being multidimensional. Of the several dimensions, the most widely investigated dimension of hearing or audition is auditory acuity, which is usually measured through pure-tone *audiometers*. Such audiometers permit the measurement of auditory acuity at different points of sound frequency, either in the group or in the individual testing situation. In testing the auditory acuity through audiometers, one ear is tested at a time and the threshold for the ascending and the descending direction of sound is determined at several frequency levels. The investigator starts with a frequency too low to be heard in the ascending direction and it is gradually increased until he perceives the sound clearly; in the descending series the examiner starts with a sound, which is clearly heard and decreased gradually until he reports no sound. Usually, the sound is given to the examinee through earphones or headphones. Recently, a test for measuring auditory acuity among children has been developed. The test is known as the *Massachusetts Hearing Test*, which requires the examinee to check Yes or No to tell whether or not he has heard the sound signal in each trial. More than 40 children can be tested through this test at a time.

Motor Dexterity Tests

Motor dexterity tests are tests which measure the co-ordination of hand, arm and/or leg movement in performing a task. There are several such tests that have been frequently used. The Crawford Small Parts Dexterity test, the Stromberg Dexterity test, the Purdue Regboard, the Benett Hand Tool Dexterity test and the Complex Co-ordination test are some of the examples of motor dexterity tests.

Artistic Aptitude Tests

Artistic aptitude tests are concerned with the measurement of artistic ability. Two types of artistic aptitudes are commonly investigated today. One type of test is exclusively concerned with the measurement of artistic appreciation and another type of test is solely concerned with the measurement of the productive ability. These two separate types of tests have been developed on the assumption that a person may have a better sense of appreciation and discrimination among several paintings but still he himself may not possess the ability to paint. One of the earliest tests of artistic appreciation was the McAdory Art Test developed by McAdory in 1929. But his test is now out of use and is remembered for its historical importance only. The Meier Art Judgement Test is one example of a test of artistic or art appreciation. The test was developed by Meier in 1929. According to Meier, artistic aptitude consists of six interrelated traits like manual skill, aesthetic intelligence, volitional perseveration, cognitive imagination, perceptual ability and aesthetic judgement. The Meier art judgement test intends to measure only the last trait of the artistic aptitude. In this test, each item consists of two very similar paintings or drawings. One of the paintings of each pair is done by a reputed painter and another one represents a slight variation of the original painting in terms of symmetry, shading, balance, use of colour, etc. The examinee is requried to choose the more appealing painting out of the two. Obviously, the test provides the extent to which the aesthetic sense of the examinee is in agreement with the opinions of art specialists. This is known as the judgement of aesthetic organization, which is considered as one of the most important factors in artistic ability. The Graves Design Judgement Test developed by Graves in 1948 and further revised in 1951 is another test of artistic appreciation. Here each item consists of two or three designs in which one is original and two are altered versions of the original. Again, the examinee is required to choose the most appealing design. Unfortunately, not much work has been done with this test and hence, it could not gain much popularity. The Horn Art Aptitude Inventory is an example of a test of artistic production. The test is a measure of the ability to produce graphic art. Hence it is an example of the work sample test. The test is solely concerned with the measurement of the ability to produce original drawing or painting. The test consists of two parts. In one part the examinee is required to make 20 drawings of very common objects (such as tree, book, etc.), and of geometric figures within a specified period of time usually in the range of 3 to 10 seconds. This is known as the scribble-and-doodle test. The second part consists of 12 rectangles, each containing a few stimulus lines. The examinee is required to make a drawing or painting in each rectangle on the basis of his imagination in a way which includes those stimulus lines. This constitutes the test of imagery. The scoring of the test is very subjective and hence, not much reliance can be placed upon this test. Other tests of artistic production are the Knauber Art Ability Test and the Lewerenz Tests in fundamental abilities of visual art.

Musical Aptitude Tests

For measuring musical aptitude, researches were carried out for nearly 40 years at the University of Iowa in the USA under the general guidelines of Seashore. As a consequence, Seashore, Lewis and Saetveit (1939) developed a test of musical aptitude, which was known as the Seashore Measures of Musical Talents. The test consists of six subtests like the pitch test, the loudness test, the rhythm test, the time test, the timbre test and the tonal memory test. The test is meant for students from grade 4 to adults. Each item of the Seashore test consists of two tones and the examinee is required to make a judgement regarding the given characteristics of the tones. For example, in the pitch test the examinee indicates whether or not the second tone is higher than the first tone and gradually each pair in the pitch test is made difficult by narrowing the difference between the first and the second tone. In the rhythm test the examinee is required to compare the rhythmic patterns of each pair, which are made either similar or different in each trial. The loudness test requires the examinee to compare the two tones in each trial and determine whether or not the second tone is stronger than the first tone. Within the time limit, the examinee is to report whether or not the second tone of each pair is longer in duration than the first tone. In the timbre test the examinee is required to discriminate between the tonal quality of the two tones in each pair, the quality of the two tones being made sometimes either equal or different. In the tonal memory test, two short series of notes (each consisting of three to five tones) are played and in the second series one of the notes is changed. The examinee is required to detect the changed note. The test does not yield a composite score for the entire test, rather total scores for each subtest are computed and each total score is converted into the percentile norms. Psychologists are of view that the test does not yield very meaningful results for a child who is below 10 and some have argued that the test is only a measure of certain types of sensory discrimination, which are essential but not sufficient for musical aptitudes (Nunnally 1970).

The Wing standardized test of musical intelligence is a battery of the musical aptitudes, which was developed by Wing (1941, 1962), a British psychologist. This test covers many of the skills required for musical training and therefore, it has been preferred to the Seashore test by most of the teachers of music. The test has seven subtests such as the chord analysis, the pitch discrimination, the tuning for pitch, the harmony, the rhythm, the intensity and the phrasing. The first three tests require sensory discrimination at a level more complex than the Seashore test and in the last four tests, the examinee is required to compare the aesthetic quality of the two versions. The first three tests may also be used as a separate short battery for younger children or for screening purposes. The entire test takes about one hour but the test may also be administered in two sessions. Norms are available for the entire battery as well as for the first three subtests separately. The reliability and the validity of the test are fairly high and satisfactory.

Test of Mechanical and Clerical Aptitudes

Although tests of mechanical and clerical aptitudes have been included in the multiple aptitude batteries, they have also been constructed as a full independent test. There are two probable reasons. First, it is not possible for any test constructor to give as much emphasis as these tests require when placed in a battery of aptitude test. Second, these two tests are more commonly used for a variety of purposes and as such, a separate status for them is essential.

A test of mechanical aptitudes is solely concerned with mechanical information, mechanical reasoning or mechanical comprehension. In some tests of mechanical aptitudes two factors, namely, the perceptual aptitudes and the spatial aptitudes, play an important role, whereas in some other tests, only items regarding the appraisal of mechanical reasoning and mechanical information are included. Examples of the former type of mechanical aptitude test are the Minnesota Paper Form Board test and the Space Relations test of the Differential Aptitude Test (DAT), and examples of the latter type of mechanical aptitude tests are the Mechanical Reasoning test of the DAT, the SRA Mechanical Aptitude test, and the Benett Mechanical Comprehension test.

Test of clerical aptitudes mainly emphasize perceptual speed and accuracy. The two well-known examples are the Clerical Speed and Accuracy Test of the DAT and the Minnesota Clerical Test. The Clerical Speed and Accuracy Test consists of two parts—each part having 100 items–which require the examinee to compare the letters and number combinations within a limited time. Each item in each part has five combinations, one of which is underlined. The examinee is to find the same one in the answersheet and draw a line under it. Part I is the practice form and Part II is the actual form, which is scored. Each correct answer is given a score of one and thus, the maximum score that an examinee can earn, is 100. The Minnesota Clerical Test also consists of two subtests, namely, the Number Comparison and the Name Comparison, each timed separately. In the Number Comparison Test, the examinee is presented with pairs of digits ranging from 3 to 12 and is required to put a mark on the blank space left between the two sets of digits. In the Name Comparison Test, proper names are given in place of the digits and the examinee is required to do the same task.

Achievement Tests

Achievement tests (also known as proficiency tests) are another important frequently used tool in the field of psychological and educational researches. Before discussing the details of the different types of achievement tests, it is essential to discuss some of their important uses. The main *uses of achievement tests are given below:*

1. Achievement tests are an effective way to check any weakness in the instructions or even slackness on the part of the examinee. If a weakness is found in the instruction, the instructor or the teacher may be asked to improve his instruction so that it may improve the subsequent instructions. Likewise, an examinee showing slackness may be motivated by special short courses so that he may be benefited from the subsequent instructions.

2. Achievement tests are also effective in the formulation of educational goals and provide a very easy means of critical examination of the content and method of instruction. A well-constructed achievement test is an effective way of producing and illustrating the major changes in any educational goal, which is often understood in terms of some commonly agreed methods,

procedures, and instrucions by which its attainment is to be assessed. Likewise, the achievement test may indicate the adequacy of the content of the syllabi. How much of the content should be retained or dropped and how much can be substituted by new ones? What are the common misunderstandings encountered in the course of content? How can the learners be motivated and made interested in bringing concrete changes in the course content? By providing answers to these questions and the like, achievement tests provide aids in the critical evaluation of the content and methods of teaching.

3. Achievement tests also help in adapting the instruction to the individual need of the learners. The performance on the achievement tests directly reveals the need for further guidance to be given to each learner and accordingly, the instruction can be modified to suit the individual need. If a learner has poor performance on the achievement test in spelling English words, it obviously indicates that the learner is in need of special guidance and course training in the spelling of English words. As such, the general instruction may be imparted in a way to suit this individual need in a satisfactory way.

Like aptitude tests, achievement tests can also be classified into two types—the general achievement batteries and the special achievement test. The achievement batteries attempt to measure the general educational achievement in several areas of the academic curriculum. They can be used from primary levels in schools to adult levels in colleges but most achievement batteries have concentrated upon the primary and the secondary levels in school. Some of the important achievement batteries developed so far are the Iowa Tests of Educational Development, the California Achievement Tests, Iowa Tests of Basic Skills and SRA Achievement Services—all of which emphasize educational skills such as arithmetical skill, spelling and reading skill and work-study skills (that is, map and graph readings). Likewise, the Sequential Test of Educational Progress (STEP) is another common example of the achievement battery, which consists of six subtests—reading comprehension, mathematics, science, social studies, listening comprehension and writing. Each subtest has multiple-choice items and is available in parallel forms. Adult Basic Learning Examination (ABLE) is also a battery of achievement tests specially meant for undereducated adults, mostly coming from economically and socially disadvantaged families. The ABLE comprises four subtests—vocabulary, reading, spelling and arithmetic. The test is meant for grades 1 through 4 (Level I) and grades 5 through 8 (Level II) and is available in parallel forms at both the levels.

Special achievement tests meant for measuring the achievements of pupils in some selected areas may conveniently be grouped into *two* distinct groups—the diagnostic tests and the standardized end-of-course examinations (Anastasi 1968). The primary purpose of diagnostic tests is to identify the educationally retarded pupils and to suggest remedial programmes for them. Such tests are available in special areas like reading skill and mathematical skill. The Stanford Diagnostic Reading Test is an example of a diagnostic test in reading skill. The test is a group test and is meant for elementary school pupils. It is available in two equivalent forms at level I, which is meant for grades 2.5 to 4.5 and at level II, which is meant for grades 4.5 to 8.5. Different dimensions of reading skill such as reading comprehension, vocabulary and word-recognition skills are measured by this test. The Durrell Analysis of Reading Difficulty is another diagnositc test in reading skill. The Durrell tests provide scores for the comprehension of oral and silent reading, rapid word recognition, listening comprehension and word analysis. The Stanford Diagnostic Arithmetic Test is an example of a diagnostic test in mathematical skill. This test is very similar to the Stanford Diagnostic Reading test in its administration. It is also a group test and is available in two equivalent forms at level I and level II. Level I is for grades 2.5 to 4.5 and level II is for grades 4.5 to 8.5. The standardized end-of-course examinations are the co-ordinated series of achievement tests for different subjects taught at either school or at college level. Since these are the co-ordinated series of achievement tests in different subjects, they provide one system of comparable norms for all tests (of the different subjects) and thus, a direct comparison of scores obtained in different subjects by the same testee is possible. Usually, the raw scores obtained on the different tests are transformed into a single scale of standard score with a fixed mean and standard deviation.

Essay-Type Tests Compared to Achievement Tests

Essay-type tests are those tests which require relatively free and/or extended written response to some questions that tend to reveal some information about the structure, dynamics and functioning of the student's knowledge as (knowledge) has been modified by a particular set of learning experiences. Such tests have been very commonly used for measuring a student's achievement in the class. When they are used for measuring the student's achievement in a classroom by a teacher, they are called teacher-made classroom achievement tests. Essay tests as measures of achievement are popular because of some potential advantages. Essay items or questions require students to communicate effectively. Here they plan their answers in such a way that there is a meaningful organisation. Moreover, essay items are not difficult to prepare if the teacher is well-versed and well-informed in the field.

But use of essay tests as measures of achievement has progressively decreased because of some inherent weaknesses like high cost of reading essay answers, difficulty of obtaining reliable and objective scoring, etc. Now essay tests are gradually being replaced by standardized achievement tests where objective items are used. The biggest merits of standardized achievement tests are that in a relatively shorter period of time, comparatively more items can be answered and the scoring can be done in a perfectly objective way.

One basic point of distinction between the essay test and achievement test is that the essay test can be used for measuring intelligence, aptitude and even personality adjustment (especially with the help of short-answer essay questions or items) apart from classroom achievement whereas achievement tests, whether in the form of essay items or objective items, can be used only for measuring proficiency acquired by the students in a field after the training. Such tests become useless in the context of intelligence, aptitude or adjustment.

Limitations of Achievement Tests

Although achievement tests are very important tests for making decisions about a pupil's performance, they have some limitations as mentioned below.

1. Scores on achievement tests cannot be taken as a basis for deciding students' promotion to the next grade. This is because most of the achievement tests are not foolproof measures of achievement in the concerned field. Many factors may determine the classroom achievement of a pupil but those factors may not be adequately assessed by the achievement test.

2. Achievement tests, as compared to intelligence tests and aptitude tests, are difficult to construct.

3. Sometimes achievement test results are taken as a measure of the teacher's effectiveness. But this is not a correct interpretation of the results because the factors that affect a pupil's achievement are numerous of which, the teacher's quality is only one.

TESTS OF CREATIVITY

Tests of creativity measure divergent thinking (Guilford 1967). Divergent thinking refers to that aspect of the mental process where the individual expands to many correct solutions of problems on the basis of a small amount of information. In this way, divergent thinking is different from convergent thinking where one seeks to synchronize or reduce several pieces of information to one correct answer. Convergent thinking is mostly measured by the general mental ability test or the intelligence test. Consistent with the above view, tests of creativity require the examinee to give the correct and appropriate response to the given situation. The following are the brief discussions of the two important tests of creativity.

1. The Torrance test of creativity was developed by Torrance (1959, 1962). This test consists of two sections—verbal and figural. The verbal section includes (1) ask-and-guess (2) product improvement and (3) unusual uses. In the *ask-and-guess* subsection, the examinee is given a picture and asked to describe what had led to the scenes in the picture and what might happen next; in the *product improvement* subsection, a picture of a toy is given to the examinee who is encouraged to suggest changes in the toy so as to make it more fun for playing, and in the *unusual uses* subsection the examinee is asked to give as many unusual uses as he or she can of a common object. The figural section requires the examinee to draw a picture representing an object and to tell an interesting and exciting story with a picture. Each subset of the figural section gives something to the examinee to start with. For example, he may be given a circle with the instruction to construct as many figures as possible, with the circle as the key part in each figure. The Torrance test can be used with both small children and adolescents. The test is administered individually and orally for children below the fourth grade. On the basis of this test, three scores are given to each examinee—*fluency* (which indicates total number of acceptable responses), *flexibility* (which indicates the number of categories in the manual used by the examinee) and *originality* (which indicates the number of responses not found in the list of frequent responses). No norms are available and therefore, each examinee's total score is evaluated against a set of common criteria for creative achievement.

2. The Remote Associates Test (RAT) is another test for creativity, which has been developed by Mednick and Mednick in 1971. The test is meant for high-school students and consists of 40 items. The items of the test are such that the examinee is presented with three words and asked to give the fourth word which may be related to each of the three words. Those three words are thought to indicate remote associative clusters and the fourth word a mediating link. One-minute time is allowed in each item and thus, the entire test takes about 40 minutes. One serious criticism of the RAT is that its validity is not established. Due to the lack of validity, Worthen and Clark (1971) contended that the RAT was a better measure of sensitivity to the structure of language rather than a measure of creative achievement.

☑ Review Questions

1. Is intelligence a combination of groups of traits or factors? Justify your answer.
2. Discuss the two-level process of intelligence. Does the global capacity view of intelligence replace the two-level-process view of intelligence?
3. Discuss the research utilities of the important types of intelligence tests in psychological researches.
4. Distinguish between aptitude tests and achievement tests. Point out some of the major limitations of achievement tests.
5. Make a comparative study of essay-type tests and achievement tests.
6. Write short notes on the following:
   (a) Intelligence quotient
   (b) Intelligence as structure of intellect
7. Make a comparative study of Wechsler scales and Binet scales.
8. Distinguish between SB4 and SB5.

◆

10

MEASUREMENT OF PERSONALITY

CHAPTER PREVIEW

- Meaning and Purpose of Personality Measurement
    - Social Traits
    - Motives
    - Personal Conceptions
    - Adjustment
- Tools of Personality Assessment
    - Self-report Inventories
    - Observational Methods
    - Projective Techniques
- Popular Strategies Involved in Construction of Personality Inventories
    - Personality Self-report Inventories
    - The Logical-Content Strategy
    - The Criterion-Group Strategy
    - The Factor Analytic Strategy
    - The Theoretical Strategy
    - Combination Strategies
    - Self-report Inventories in India
- Overcoming Distortions in Self-report Inventories
    - Establishment of Rapport
    - Use of Forced-Choice Technique
    - Concealing the Main Purpose of the Test
    - Use of Verification and Correction Keys
- Situational Tests
- Measurement of Interests, Values and Attitudes
    - Meaning and Types of Interest Test
    - Value Tests
    - Measures of Attitudes

MEANING AND PURPOSE OF PERSONALITY MEASUREMENT

The term 'Personality' refers to the total functions of an individual who interacts with his environment. Such a definition automatically includes all traits as the main themes of the personality. The purpose of the measurement of personality is to describe a person in terms of traits. Trait is nothing but the observed consistency of behaviour of a person. Traits are not directly observed but inferred from the consistent behaviour of an individual. The most general cues to traits are what the person does, how he does it and how well he does it. Nunnally (1978, 354) has opined that personality measurement aims at studying the following four broad types of traits:

Social Traits

Social traits are those traits which determine how persons interact with other persons in the society. Examples of typical social traits are friendliness, honesty, dominance, responsibility, shyness, etc. Thus, within social traits are included traits related to temperament and character.

Motives

Motives here include the nonbiological drives such as need to earn money and prestige, need for academic achievement, need for affiliation, aggression, etc. These nonbiological needs are often said to constitute what is known as *personality dynamics*.

Personal Conceptions

Under the trait of personal conceptions are included those methods which determine people's attitude toward the self and others, a person's values, interests, etc.

Adjustment

Adjustment includes traits like the freedom from emotional worries or instability and other disruptive behaviour. Maladjustment is the opposite of adjustment and includes pathological traits like hallucinations, hysteria, imaginary illness, psychoses, etc.

One general characteristic of these four types of traits is that they are correlated with each other, that is, they are not independent. For example, a social trait, say dominance, is likely to influence motives, personal conceptions (that is, interests, attitudes, etc.) and adjustment. Likewise, an individual's motive is likely to influence his interaction with others in society (that is, social traits), his personal conception and his adjustment, too. If an individual is highly prejudiced against a particular caste (an example of personal conception), his social interaction, motives and adjustment would all be similarly affected. Likewise, if a person has a satisfactory general adjustment, his social traits, motives and personal conception would be in congruence with social norms. On the other hand, if an individual is maladjusted, he would have extreme social traits, eccentric motives and personal conceptions.

Personality measurement aims at measuring the above four broad categories of general traits. For the measurement of social traits, motives and adjustment, more or less a similar set of principles is found but to measure personal conception, which includes attitude, interest and values, a different set of principles is necessary.

TOOLS OF PERSONALITY ASSESSMENT

There are three most common tools (or methods) of personality assessment.

1. Self-report personality inventories, or Structured personality tests
2. Observational methods including situational test
3. Projective techniques

A brief discussion of each of these is given below.

Self-report Inventories

Self-report inventories, also known as personality inventories, are the self-rating questionnaires, where the individual describes his own feelings, environment, and reactions of others towards himself. In a nutshell, on the self-report inventories a person reports about himself in the light of the questions (or items) put therein. Hence, the method is known as a self-report inventory. Self-report inventories are further classified into the following five types:

1. Inventories that attempt to measure social and certain other specified traits such as self-confidence, dominance, ego-strength, extroversion, responsibility, etc. The Bernreuter

Personality Inventory, the Eysenck Personality Questionnaire and the Differential Personality inventory are some of the examples of this category.

2. Inventories that attempt to evaluate the adjustment of persons to different aspects of the environment, such as school, home, health, etc. The Bell Adjustment Inventory is the best example.

3. Inventories that attempt to evaluate pathological traits such as hysteria, paranoia, hypomania, depression, schizophrenia, etc. The Minnesota Multiphasic Personality Inventory (MMPI) is the best example.

4. Inventories that attempt to screen individuals into two or three groups. The Cornell Index is the best example of such an inventory. It screens the persons into two groups—those having psychosomatic difficulties like asthma, peptic ulcers, migraine, convulsive disorders, etc., and those not having them, that is, those who are normal.

5. Inventories that attempt to measure attitudes, interests and values of persons. The Kuder inventories (vocational, occupational and personal), the Strong Vocational Interest Blank, the Allport–Vernon Study of Values are some of the best examples of this category of self-report inventories.

It does not follow, however, that the above five classifications of inventories have nothing in common. In reality, the classification is based upon the purpose and the nature of item content. All the above self-report inventories are based upon the same principle, which states that behaviours are nothing but the manifestation of trait and one can find out the presence or absence of a trait by means of assessing the behaviour.

A review of some of the representative personality inventories has been done later in this chapter.

Observational Methods

Observational methods are distinct from self-report inventories. Observational methods provide either a structured or unstructured situation. A structured situation is a controlled situation whereas an unstructured situation is an uncontrolled situation. Persons whose personality traits are to be observed are put in either of these two situations and careful, impartial and accurate observations are made by the observers. The observation becomes the basis for assessing the personality traits. In some observational methods, however, a departure is made from the above set procedure. The differences in observations made by different observers reflect the subjectivity in the observations of the observers. A detailed discussion of the observation method appears in Chapter 12.

Projective Techniques

Projective techniques are the most popular method of assessing the personality of the individual. As its name implies, the person whose traits are to be studied is asked to describe an unstructured stimulus or situation and through his responses, his needs, drives, motives, fears, etc., are revealed. (For a complete discussion of projective techniques, refer to Chapter 11.)

POPULAR STRATEGIES INVOLVED IN CONSTRUCTION OF PERSONALITY INVENTORIES

Personality Self-report Inventories

Self-report personality inventories (also technically called *Structured Personality tests*) are the most popular measure of personality. In fact, structured personality tests try to evaluate personality traits, personality types, personality states and some other related aspects including self-concept. Briefly, *personality traits* refer to a relative disposition to think, feel and act in a consistent manner. *Personality types* refer to a general description of people. Personality states refer to emotional reactions that vary from one situation to another. *Self-concept* refers to a relatively organized and consistent set of assumptions that a person has about himself or herself.

In a structured personality test, the subject or testee is asked to respond to written statements marking them in some way as 'True' or 'False' to indicate whether the statement applied to them. That is the reason why it is called *structured* or *objective method* of personality assessment as distinguished from the projective method. Here each statement is structured and lacks ammbiguity, whereas in projective method of personality the stimulus is ambiguous and the testee has no definite guidelines about what type of response is needed.

A review of literature reveals that the construction of structured personality tests is based upon certain popular strategies. There are five such strategies: *the logical-content strategy, the criterion-group strategy, the factor-analytic strategy, the theoretical strategy* and *the mixed strategy.* We shall discuss each of these strategies separately.

The Logical-Content Strategy

As its name implies, in this strategy the test constructor uses various types of reason and deductive logic in the development of the test. This strategy assumes that the test item describes the person's personality and behaviour. For example, if the testee or subject marks 'True' for the statement "I like to participate in social activities" then testers assume that he or she really likes social participation. Thus the test constructor assumed the face validity of a test response. Initial efforts to assess personality used this logical-content approach as the primary approach.

Specific tests of personality were not developed until World War I when psychologists realized the need to distinguish people on the basis of emotional well-being. To meet this need, the first personality inventory was developed by Woodworth (1920) and it was named *Woodworth Personal Data Sheet,* whose purpose was to identify military recruits likely to develop emotional problems in combat. It consisted of 116 questions to which the testee responded either 'Yes' or 'No'. The items of the test had been selected from known symptoms of emotional disorders and from the questions asked by clinical psychologists and psychiatrists in their screening interviews. The test produced a single overall score providing a global measure of the personality functioning.

The success of the Woodworth Personal Data Sheet in solving the problem of mass screening attracted several other psychologists who devoted attention to the development of the logical-content test of personality. Two of the best-known early tests based on logical-content strategy were the *Bell Adjustment Inventory* and the *Bernreuter Personality Inventory.* The Bell Adjustment Inventory assessed adjustment of persons in various areas such as home life, social life, health life and emotional life. The Bernreuter Personality Inventory measured six personality traits like sociability, confidence and introversion. Both these tests were originally published in the 1930s and yielded not one but more than one overall score. In fact, this laid the foundation of many such modern tests of personality that yielded multiple scores.

Although the logical-content approach was very popular during the initial phase of personality test construction, its weaknesses soon became evident. These weaknesses were mainly related to its assumptions side. The logical-content approach assumed that the subject complies with the given instructions and provides an honest response; the subject goes through each item and is also capable of evaluating his or her own behaviour; the subject, test constructor and test interpreter all define the items of the test in a similar manner. A wide variety of researches questioned the validity of all these assumptions. In fact, none of these assumptions was necessarily true. Structured Personality tests based on the logic of face validity were so sharply criticised that such tests were more or less fully discarded (Ellis 1946; Landis 1936; Mc Nemar & Landis 1935). As a consequence, a new strategy called the criterion-group strategy emerged as a rescue of new conceptualization in personality test construction.

The Criterion-Group Strategy

The criterion-group strategy is another method of personality test construction with its enthusiasm to make an improvement over the logical-content strategy. This is in fact an empirical approach that relies on data collection and various types of statistical analyses for determining the meaning of a test response or nature of personality (Kaplan & Saccuzzo 2001). In this strategy of personality test construction, the test constructor selects two groups of people: a *criterion group* and a *control group*. The criterion group consists of persons or individuals who share certain characteristics such as hysteria or schizophrenia. The control group consists of persons that represents the general population. He then selects and administers a group of items that make a distinction between the criterion group and the control group. The items that distinguish these two groups are selected in the test. Thus the actual content or face validity of an item in the criterion-group strategy is of little value. Subsequently, such items are cross-validated by checking how well it distinguishes an independent criterion group consisting of individuals also known to possess the characteristics to be measured, from a control group. If the scale is found to distinguish these two groups then it is said to have been cross-validated. Subsequently, data from normal control groups are used to obtain standard scores. Thus one can easily determine how far above and below the mean of the normal group each new subject scores in terms of the standardized units. After the test has been constructed and cross-validated, some additional research is conducted to ascertain empirically what it means when the subjects endorse many items of a particular dimension of the test.

Here we shall discuss two well-known tests which have been developed using criterion-group strategy: *Minnesota Multiphasic Personality Inventory* and *California Psychological Inventory.*

Minnesota Multiphasic Personality Inventory (MMPI)

The Minnesota Multiphasic Personality Inventory (MMPI) is a true–false self-report, or questionnaire measure of personality. This inventory was developed in the early 1940s by Hathaway & McKinley (1943). It is a very important means for detecting disabling psychological abnormalities. It is called multiphasic because it was designed to detect several psychiatric and pathological problems. Originally, this inventory had 550 items. Each item of this original inventory had three options: 'True', 'False', 'Cannot say'. There were ten clinical scales and four validity scales. The names of the clinical scales are:

1. HS: Hypochondriasis
2. D: Depression
3. Hy: Hysteria
4. Pd: Psychopathic deviate
5. Mf: Masculinity–femininity
6. Pa: Paranoia
7. Pt: Psychosthenia
8. Sc: Schizophrenia
9. Ma: Mania
10. Si: Social Introversion

The clinical scales were designed to identify psychological disorders like hysteria, depression, etc. The validity scales intended to measure test-taking aptitude and to assess whether the subject was demonstrating a normal, honest approach to the test. The four validity scales are: Lie scale (L), Infrequency Scale (F), Correction Scale (K) and Cannot Say (or ?). Although these scales are named as validity scales, they are not concerned with validity in the technical sense. In reality, these validity scales represent checks on carelessness, misunderstanding, malingering and other different kinds of response sets and test-taking aptitudes. In particular, the L scale was designed to detect such subjects who tried to present themselves in an overly favourable way. The F scale consists of such items that are scored infrequently (less than 10%) by the normal population. High scores on the F scale invalidates the profile (Shores & Carstairs 1998). The purpose of this scale is to detect such individuals who attempted to fake bad. K scale served the same purpose as that of L scale but it had been empirically constructed. A high score on the K scale may indicate defensiveness or an attempt to 'fake good', and a low K score tends to represent excessive frankness and self-criticism or a deliberate attempt to 'fake bad'. They cannot say that the score represents the number of items omitted or double marked. Omission up to 10 items appears to have little effect on the overall test results but if this count exceeds 30 items, the test record is considered highly suspect and probably becomes invalid (Anastasi & Urbina 1997). Thus, unlike the other three validity scales, it has no items of its own.

The subjects obtained a raw score on each of the 10 clinical scales based on the number of items they have marked in the scored direction. Subsequently, raw scores are converted into McCall's T scores with a mean of 50 and a standard deviation of 10.

Recently, the original MMPI has been revised and reconstructed into two separate versions: the MMPI-2 (Butcher et al. 1989) and the MMPI-Adolescent (MMPI-A) (Butcher et al. 1992).

The MMPI-2 consists of 567 affirmative statements which are responded by the subject as either 'True' or 'False'. The first 370 items, which are more or less identical to MMPI except for editorial changes, provide the responses needed to score the original 10 clinical and 4 validity scales. Thus, all the 10 clinical and four validity scales have been retained in MMPI-2. The remaining 197 items (of which 107 are new) constitute the full complement of 104 new, revised content, and supplementary scales and subscales that make up the complete inventory. The basic profile form for the MMPI-2 includes 10 clinical and four validity scales taken over from the original version. Apart from this, there are separate profile forms for 15 content scales, 27 content component scales, 21 supplementary scales and 28 Harris–Lingoes subscales. All these scales of MMPI-2 are scored by using the MMPI-2 normative sample of 2600 adults having an age range of 16 to 84 years. Some important examples of content scales are HEA (health concerns), TPA which evaluates the hard-working, irritable, impatient Type-A personality, FAM (family problems) which evaluates family disorders and possible child abuse and WRK (work interference) that examines the behaviours and attitudes likely to interfere with work performance. Examples of supplementary scales are Anxiety, Repression, Ego-Strength, and Mac Andrew Alcoholism Scale-Revised.

One important feature of the MMPI-2 is the inclusion of three new validity scales that can help to assess the care and veracity with which the subject responds to the inventory. They are *Back F* (F_b) scale, *the Variable Response Inconsistency Scale* (VRIN) and the *True Response Inconsistency Scale* (TRIN). The F_b scale is basically an extension of the original F scale for items that appear in the second half of the inventory and provides a check on validity and correction throughout the test. The VRIN scale attempts to evaluate random responding and the TRIN scale attempts to assess acquiescence tendency, that is, tendency to agree or mark 'True' regardless of the content. The VRIN consists of matched pairs of items that have similar content. When the subject marks the pairs in the opposite direction, he earns a score on the scale. The TRIN Scale also consists of matching pairs of items with opposite contents. For example, to receive a point on TRIN Scale, the subject should mark 'True' to both items like '*I feel disturbed most of the time*' and '*I remain relaxed most of the time*'.

Finally, it must be made clear that MMPI-2 is scored for 10 clinical scales in addition to its validity scales. With the exception of one clinical scale, that is, 'social introversion', the other nine clinical scales were constucted in the usual criterion-keyed manner in which the responses of clinical subjects and normal subjects were contrasted. Social Introversion was, however, developed by contrasting the responses of college students high and low in social introversion.

The psychometric properties of the MMPI and MMPI-2 are comparable (Gaston et al. 1994). Median test-retest coefficients range from the low .50s to .90s. Likewise, median split-half reliability coefficients for both MMPI and MMPI-2 run in the .70s. Despite satisfactory reliability coefficients, MMPI-2 has some problems. Importantly, such problems are that its items overlap among the scales, an imbalance shown in true–false keying, high intercorrelation among scales (for example, scales 7 and 8 correlate between .64 and .87) and lack of generalizability across demographic variables (Kaplan & Saccuzzo 2001).

Despite these limitations, the objective scoring of MMPI and MMPI-2 made it popular both as a research and clinical tool. According to Graham (1993), it has been used in more than 10,000 studies. One survey has revealed that it is preferred by about 90% of the clinical psychologists (Aiken 1988).

MMPI-A is the new form of MMPI developed specially for use with adolescents. It consists of 478 items covering areas specifically relevant to adolescents such as school and family problems and above all, the provision of age-appropriate norms. MMPI-A basically retains most of the features of the MMPI and MMPI-2 including 10 clinical scales and four validity scales. The age range of the standardization sample was 14 to 18 years.

Besides, MMPI-A has its own validity scales (F_1 and F_2) as well as some content scales, supplementary scales and subscales that are unique to it and some that are similar to MMPI-2.

In Conclusion, it can be said that both MMPI-2 and MMPI-A are today very promising and important tools of assessing personality, and publications of books and articles on these two seem to be continuing unabated (Pope, Butcher & Seelen 1993; Butcher, Graham & Ben-Povath 1995).

California Psychological Inventory

The California Psychological Inventory (CPI) was originally developed by Harrison Gough in 1957 and it, then, consisted of 480 items. The inventory has undergone revision, and as a consequence, its original length of 480 items has been shortened to 462 items in the 1987 revision and most recently to 434 items which is the third edition (Gough & Bradley 1996). All the 434 items are to be answered as either 'True' or 'False' and yield scores on 20 scales.

CPI is, in fact, a structured personality test constructed primarily by the criterion-group strategy. One of the major features of CPI is that more than a third of the 434 items are drawn from MMPI and it is specifically used for assessing personalities of normal adult populations. On this point, CPI differs from MMPI which assesses only the personalities of abnormal persons. In the latest version of CPI, of 20 scales, three are validity scales designed to assess test-taking attitudes. The three validity scales are: *Well-being* (Wb), *Good impression* (Gi) and *Communality (Cm)*. The Wb scale is based on responses by normals asked to 'fake bad'; the Gi scale is based upon responses by normals asked to 'fake good' and the Cm scale is based on a frequency count of highly popular responses. In this way Wb, Gi and Cm are constructed, respectively, to detect subjects who 'fake good', 'fake bad' and respond randomly. The remaining 17 scales provide scores on personality dimensions like Dominance (Do), Sociability (Sy), Capacity for Status (Cs), Social Presence (Sp), Self-acceptance (Sa), Independence (In), Empathy (Em), Responsibility (Re), Socialization (Sc), Self-control (Sc), Tolerance (To), Achievement via conformance (Ac), Achievement via Independence (Ai), Intellectual Efficiency (Ie), Psychological-Mindedness (Py), Flexibility (Fx) and Feminity (Fe).

Like MMPI-2, all CPI scores are converted into standard score scales with a mean of 50 and a standard deviation of 10. The advantage of CPI is that it can be used for assessment of normal personalities. The MMPI and MMPI-2 generally don't apply to assessment of normal subjects. Therefore, the future of the CPI as a measure of normal personalities has good potentiality (Bolton 1992; Groth-Marnat 1999). Cross-cultured studies have shown that the CPI is useful in explaining personality differences between ethnic groups (Dana 1993; Davis, Hoffman & Nelson 1990).

Researches have shown that CPI is very useful for predicting the following:

(i) high school and college achievement
(ii) grade point average in medical institutions
(iii) effectiveness of student and teachers
(iv) effectiveness of police and military personnels
(v) leadership and executive successes

It has also been found that this test is useful in identifying adolescents or adults who follow a delinquent or criminal path in their lives.

Reliability coefficients of CPI are very similar to those reported for the MMPI. Short-term test-retest reliability coefficients range from .49 to .90 and long-term coefficients range from .38 to .77. However, CPI shares some of the limitations of MMPI. Like the MMPI, the various scales of CPI have shown considerable intercorrelation. Similarly, like the MMPI, true–false scale keying is often extremely unbalanced.

The Factor Analytic Strategy

The factor analytic strategy uses the technique of factor analysis to derive empirically the major dimensions of personality. In fact, factor analysis is a statistical technique for reducing the redundancy in a set of intercorrelated scores. The principal-components method is one such method of factor analysis (Hotelling 1933) that tends to find the minimum number of common factors that can account for an interrelated set of scores. Factor analysis starts with a larger database consisting of the intercorrelation of a larger number of items or tests. Subsequently, these intercorrelations are factor analyzed to find the minimum number of factors that acount for as much of variability in the data as possible. Then, they try to label these factors by finding out what the items related to a particular factor have in general.

Guilford, Cattell and Eysenck were the three very important psychologists who tried to develop personality test using the factor analytic strategy. Guilford made the pioneer effort by determining intercorrelations of a wide variety of tests and then factor analyzed the results in an attempt to find the main dimensions underlying all such tests. After locating such a factor, items that correlated high with these factors could be finally taken in the test to capture the major dimension of personality. The result of this initial attempt using factor analytic strategy was a series of inventories, which were published by Guilford and his associates in the 1940s. For example, Guilford (1940) found five factors after factor analyzing the test of introversion–extroversion. These factors were: *social introversion, thinking introversion, depression, cycloid tendencies* and *restraint*, popularly known as the *Inventory of factors STDCR*. Further factorial studies by Guilford and Martin led to the development of two other personality inventories, Guilford–Martin Inventory factors GAMIN (1943a) and Guilford–Martin Personality Inventory (1943b). After combining two highly correlated factors to avoid confusion and redefining the remaining other factors, a single ten-factors inventory was developed. This inventory is known as *Guilford–Zimmerman Temperament Survey* (Guilford & Zimmerman 1956). The ten factors or dimensions were: General activity (G), Restraint (R), Ascendence (A), Sociability (S), Emotional stability (E), Objectivity (O), Friendliness (F), Thoughtfulness (T), Personal relations (P) and Masculinity (M). Each of these ten dimensions is measured by 30 items. All 300 items are self-statements as in MMPI and MMPI-2, and the subject must indicate 'Yes' or 'No' for each statement like MMPI. In the survey, three verification keys are included to detect falsification and for evaluating the validity of the profile.

However, this first factor analytic structured personality test failed to be popular partly because it was overshadowed by MMPI and partly because of its arbitrary subjective way of naming the various factors. The current position is that Guilford–Zimmerman Temperament Survey mainly serves only historical interests.

Cattell's Personality Questionnaire

Cattell is another important psychologist who widely used this factor analytic strategy in developing a structured personality test. On the basis of their factor analytic strategy, Cattell and his co-workers have developed a number of personality inventories. Of these, the best known is the *Sixteen Personality Factor Questionnaire*, currently being run in its fifth edition (R B Cattell, A K Cattell, H E Cattell 1993; Russell & Karol 1994). The 16 PF was originally published in 1949 and is meant for the assessment of personality for ages 16 and over. It is a very widely used forced-choice test which is available in five separate forms. Each form consists of declarative stems that force the testee to respond to a specific situation by choosing from among two (Form E) or three forced-choice options (Forms A, B, C and D). It yields 16 scores separately on 16 traits of personality, each of which is bipolar. Each of these 16 factors or source traits has been given a letter-symbol or letter–number combination. All these 16 factors and their letter symbols have been presented in Table 10.1

Table 10.1 Cattell's 16 personality factor and their letter symbol

Low	Factors	High	Letter Symbol
Reserved	*Warmth*	Outgoing	A
Less Intelligent	*Intelligence*	More Intelligent	B
Stable Ego Strength	*Ego-Strength*	Emotionality/Neuroticism	C
Humble	*Dominance*	Assertive	E
Sober	*Impulsivity*	Happy-go-lucky	F
Expedient	*Conformity*	Conscientious	G
Shy	*Boldness*	Venturesome	H
Tough minded	*Sensitivity*	Tender minded	I
Trusting	*Suspiciousness*	Suspicious	L
Practical	*Imagination*	Imaginative	M
Forthright	*Shrewdness*	Shrewd	N
Placid	*Insecurity*	Apprehensive	O
Conservative	*Radicalism*	Experimenting	Q_1
Group-tied	*Self-Sufficiency*	Self-Sufficient	Q_2
Casual	*Self-discipline*	Controlled	Q_3
Relaxed	*Tension*	Tensed	Q_4

This fifth edition of the 16PF is available in only one form and contains 185 items, most of which have been selected from previous forms of the questionnaire on the basis of their sound psychometric properties (Anastasi & Urbina 1997). Consistent with the factor analysis strategy, items that correlated highly with each of the 16 major factors or source traits, were included and those with relatively low correlations were discarded from the test.

One of the unique features of the 16 PF is the inclusion of '*Problem Solving questions*' which contains 15 items presented contiguously at the end of the inventory. These items comprise the Reasoning scale which is basically intended to act as a quick measure of mental ability. Besides these, the questionnaire now also includes measures for detecting three indices of response style, namely, measures for assessing random responding, acquiescence tendency and attempts to show socially desirable or undersirable qualities.

A brief description of each of the 16 factor scales is presented below.

1. *Reserved–outgoing, Scale A:* Scale A, in fact, is theoretically a measure of warmth. Persons securing low scores (Sten score 1–3) on this scale generally seem to possess negative traits like aloof, formal, showing, hostility and suspiciousness. These traits are associated with paranoid schizophrenia, although the scale does not measure such a pathological dimension. Such persons find contact with other persons to be anxiety provoking or even painful. On the other hand, persons scoring high on Scale A (Sten score 8–10) like to be around other people. Such persons are more emotionally expressive and they may feel disturbed if they are forced into a situation where they are not allowed contact with others. Such persons are extrovert and easygoing.

2. *Less intelligent–More intelligent, Scale B:* Scale B measures intelligence. Cattell distinguishes between two kinds of intelligence—crystallized intelligence and fluid intelligence. Fluid intelligence refers to innate abilities, whereas crystallized intelligence refers to the abilities that are products of the cultural learning, training and schooling. Low scorers on Scale B tend to show low mental capacity, poor judgement, less perseverance, poor ability to work with abstract problems and low morales. Poor scores on this scale indicate anxiety, inability to concentrate and poor intelligence. High scores on scale B indicate high intelligence, ability to work with abstract ideas, good morale, good judgement and high perseverance. It may invariably be mentioned here that Cattell has also separately developed *Cattell Fair Intelligence Test* which is designed to assess fluid or innate intelligence, independent of cultural learning and training.

3. *Stable Ego-Strength—Emotionality/Neuroticism, Scale C:* Scale C measures ego strength of the person. The essence of this factor is inablity to control one's impulses or to deal realistically with various problems (Cattell 1965). High scorers on this scale tend to show emotional maturity, general lack of anxiety and ability to deal with frustrating or very difficult situations. Low scorers on this scale demonstrate a general emotional liability, inability to handle frustration, evasion of responsibility and tendency to show excessive worry over trivial things. From this description, it is clear that scale C plays an important role in identifying many psychiatric disorders.

4. *Humble–Assertive, Scale E:* Scale E is a measure of dominance. Low scorers tend to be unsure, retiring, modest, meek, quiet and obedient. Therefore, they are associated with conventionality, docility, dependence, submissiveness and obedience. High scorers are self-assertive, boastful, conceited, aggressive, forceful, vigorous and egotistical. Therefore they are associated with aggressiveness, hostility, dominance, rebelliousness and independence. A high E score is not necessarily pathological but can be a serious sign in connection with other scale elevations that obviously suggest a psychopathic personality or a tendency towards emotional outbursts. Cattell (1965) has shown that men and boys tend to score higher on dominance than women and girls. Likewise, neurotics who are gradually improving show increase in dominance (Ryckman 1993).

5. *Sober–Happy-go-lucky, Scale F:* Scale F measures impulsivity. Happy-go-lucky persons or high-surgency people (high F scorers) are characterised by being joyous, cheerful, sociable, responsive, energetic, witty, humorous, as well as talkative. Sober people or people with desurgency (low F scorer) are characterized as depressed, pessimistic, seclusive, introspective, worrying and retiring. Cattel has described low F persons as dull, rigid, anxious and depressed. Such individuals are frequently associated with depressive states. On the other hand, high F scorers reflect impulsivity, extroversion and adaptability (Karson & O'Dell 1976).
6. *Expedient–Conscientious, Scale G:* Scale G is a measure of group conformity and items of this scale are very similar to MMPI L scale. High scorers are associated with high superego strength, responsibility, conscientiousness, moral correctness and dutifulness. Such persons are either excessively rigid and moralistic or are answering in a 'faking good' condition. Low scorers on scale G tend to be self-indulgent, undependable, frivolous and fickle and are generally unconcerned about group norms and standards.
7. *Shy–Venturesome, Scale H:* Scale H measures boldness. High H scores are associated with extroversion, social boldness, responsiveness, interest in opposite sex and insensitivity to danger signals. On the other hand, low H scores are associated with restraint, shyness, emotional cautiousness and unfriendlines. Cattell has suggested that low H scores are associated with a tendency towards schizoid personality.
8. *Tough-minded–Tender-minded, Scale I:* Scale I is a measure of emotional sensitivity. High scores on Scale I are associated with dependence, imaginativeness, attention-seeking behaviour, emotional sensitivity, insecurity and hypochondriasis. On the other hand, low scores are associated with self-reliance, lack of sentimentality, cynicism, logic, practicality and lack of hypochondriasis. Some studies demonstrate a close relationship existing between Scale I and MF scale of MMPI.
9. *Trusting–Suspicious, Scale L:* Scale L measures suspiciousness. High L scores are associated with being jealous, dogmatic, frustrated, domineering, and irritable. In extreme scale, such high scores may show sign of paranoia. Low scores on this scale are characterized as being permissive, trusting, tolerant, nonhostile, trusting and uncritical.
10. *Practical–Imaginative, Scale M:* Scale M measures imagination. Persons having low scores on this scale are described as being practical, objective, conservative, conventional and not over-imaginative. On the other hand, persons having high scores are characterized by being fanciful, subjective, unconventional, imaginative and interested in various types of art and philosophy. High scorers tend to do better in jobs requiring imagination and artistic skills. However, individuals with low score on the M scale tend to do better on jobs with practical and realistic demands.
11. *Forthright–Shrewd, Scale N:* Scale N is a measure of shrewdness. This scale is found to correlate significantly with Hysteria Scale of MMPI. Persons having high scores on this scale are characterized by worldliness, ambitious, social awareness, calculating mind, and having a strong tendency to take advantage of situations. Persons having low scores on this scale are characterized by spontaneity, vagueness, lack of self-insight, passivity, genuineness and a blind trust in human nature. According to Karson and O'Dell (1976), the scale N is the least useful scale of the 16 PF.
12. *Placid–Apprehensive, Scale O:* This scale measures guilt proneness. High O scorers are self-confident, resilient and placid. Low O scores are suggestive of an uncontrolled and adequate individual who is cheerful and is likely to act when necessary. Such low scores may be associated with weak superego controls. On the other hand, high O scorers are

persons who are insecure, guilt-prone, apprehensive and troubled individuals who are more likely to be anxious, depressed and very much sensitive to approval or disapproval of others. From such a description, it is obvious that this scale plays an important role in clinical analysis of the 16PF.

13. *Conservative–Experimenting, Scale* Q_1: Scale Q_1 is a measure of rebelliousness. High Q_1 scorers are associated with radicalism and liberalism. Such persons have a tendency to be experimenting, analytical and free thinking. On the other hand, low Q_1 scorers are associated with conservatism. Such individuals tend to show respect for traditional things and are unwilling to change the way things are done. Persons having extreme Q_1 scores are unable to accept authority, a trait often found in sociopathic personalities.
14. *Group-tied–Self-sufficiency, Scale* Q_2: This scale measures self-sufficiency. High scorers are characterized by being self-sufficient persons who like to depend much upon their own resources and judgements. Low scorers are persons who are group dependent, frequent joiners and also good followers.
15. *Casual–Controlled, Scale* Q_3: Scale Q_3 is a measure of the ability to bind anxiety. High scorers are persons who are controlled, socially precise and compulsive. Their willpower is also strong. Such persons think carefully before they act and do not let emotions disturb routine. These persons are well-controlled and dependable. In a nutshell, such persons have sound mental health. On the other hand, low scorers are persons who are characterized by a lack of control, tendency to show disrespect and carelessness to social rules and tendency to follow one's own urges. Such persons also possess the trait of overreactivity and they often fail to handle stress in a productive way. If a low Q_3 is found in the presence of some other anxiety indicators, it is enough to suspect that the person is having some kind of emotional trouble.
16. *Relaxed–tensed, Scale* Q_4: This scale measures free-floating anxiety and tension. High scorers on Q_4 tend to exhibit characteristics like tension, frustration and a highly anxious approach to life. Such scores indicate a person with extreme problems, a cry for help and showing socially undesirable tendencies. This scale, in fact, is the best indicator of neurotic anxiety. Low scorers, on the other hand, tend to exhibit low tension, low anxiety, and a relaxed approach to life. Such persons don't exhibit frustration even in a situation that may rightly produce this.

Test-retest reliability and internal consistency reliability for the 16 primary factor scales of the 16 PF are better. Short-term test-retest reliability coefficient for the 16 source traits range from .65 to .93 with a median coefficient of .83. Despite the scientific method used for deriving the factors, 16 source traits of the 16 PF do intercorrelate with each other, and sometimes this correlation becomes as high as .75 (Cattell, Eber & Tatsuoka 1970). For dealing with such overlapping among the factors, 16 factors themselves were factor analyzed that yielded four second-order factors for which scores can be obtained (Kaplan & Saccuzzo 2001).

From the 16 PF scale, some parallel inventory for ages from 12 to 18 and for ages from 8 to 12 have been constructed. The former is known as the *Junior Senior High School Personality Questionnaire* and the latter is called the *Children's Personality Questionnaire*. For extending the 16 PF scale to the assessment of psychopathology, items relating to psychological disorders have been factor analyzed, resulting in 12 new factors in addition to 16 factors which are needed to assess normal personality. These new factors together constituted a new clinical instrument called the *Clinical Analysis Questionnaire* or CAQ (Delhees & Cattell 1971). The twelve abnormal source traits assessed by the CAQ are presented in Table 10.2.

Table 10.2 Factors of CAQ

Factors	Traits
D_1	Hypochondriasis
D_2	Zestfulness/Suicidal disgust
D_3	Brooding discontent
D_4	Anxious depression
D_5	Energy euphoria
D_6	Guilt and resentment
D_7	Bored depression
P_a	Paranoia
P_P	Psychopathic deviation
S_C	Schizophrenia
A_S	Psychasthenia
P_S	General psychosis

Despite the best care, when 16 PF scale is compared with MMPI and MMPI-2, it is found to be less useful. Besides this, the claim of the 16 PF to measure the basic source traits of personality are also not much appreciating. Despite all these limitations, the 16 PF remains one of the exemplary illustrations of factor analytic approach to personality testing.

Eysenck Personality Questionnaire (EPQ)

Based on a lifelong programme of factor analytic questionnaire, Eysenck and Eysenck (1975) developed a series of tests designed to measure normal and abnormal dimensions of personality. Eysenck identified three major dimensions of personality: Psychoticism (P), Extraversion (E) and Neuroticism (N). The Eysenck Personality Questionnaire (EPQ) comprises items that intend to measure these three dimensions of personality. The EPQ consists of 90 statements to be answered in terms of either 'Yes' or 'No' and is specially suited for persons aged 16 and older. It also incorporates a Lie (L) scale to assess the validity of the testee or examinee's responses. Also, a junior EPQ is available for assessing these dimensions among children aged 7 to 15 and it consists of 81 statements. A brief description of these three scales are as under:

(1) P Scale: P scale assesses the dimension of psychoticism which is not equivalent to psychosis such as schizophrenia, although a schizophrenic is expected to score high on P scale. It assesses traits like poor concentration, poor memory, insensitivity, liking for unusual things, disregard for danger and convention, cruelty, lack of caring for others. Such persons are considered peculiar by others. A high score on P scale indicates impulsivity, aggressive and hostile traits, empathy defects and a preference for liking odd or unusual things. Antisocial personality and schizoid personality often obtain high scores on this dimension. A low score on P scale indicates some derivable characteristics like empathy and interpersonal sensitivity. A few examples of items of P scale are:

Do you take risk just for fun? (T)

Do you often break the rule? (T)

(2) E Scale: E scale assesses the dimension of extraversion and its polar opposite introversion. High scores on E scale indicates tendency to be outgoing, preference for activities involving contact with other people, desire for novelty. Such persons are fun-loving and

gregarious. Low scores on this scale indicate introverted traits such as preference for solitude and quiet activities. Such persons show tender mindedness, introspectiveness and seriousness. A few examples of items of E scale are:

Do you like plenty of excitement? (T)

Are you quiet when with others? (F)

(3) N Scale: N scale assesses the dimension of neuroticism that includes traits like slowness in thoughts and actions, suggestibility, tendency to repress unpleasant facts, lack of sociability, below-average emotional control, willpower and capacity to exert self. A high score indicates that the person is nervous, maladjusted and overemotional and a low score indicates that the person is stable and confident. A few examples of items of N scale are:

Are your feelings easily hurt? (T)

Do you feel dullness in life? (T)

A major focus of research with the EPQ has been to find out the empirical correlates of extraversion and its opposite introversion, and such researches have linked several perceptual and physiological factors to the dimension E-I. Some of the important such linkages are:

(i) Extroverts have a greater need for entertaining external stimulation.
(ii) Extroverts are readily conditioned to stimuli associated with sexual arousal.
(iii) Extroverts are more suggestible than introverts.
(iv) Introverts are vigilant in watch keeping.
(v) Introverts' performance on signal detection tasks are comparatively more improved.
(vi) Introverts are less tolerant of pain but more tolerant of sensory deprivation.

The psychometric properties of the EPQ are satisfactory. The one-month test-retest reliabilities were .78 (P), .89 (E), .86 (N) and .84 (L). The internal consistency reliabilities were in the .70s for P and the .80s for the remaining three scales. The construct validity of EPQ is also well established in several studies using emotional, behavioural, attentional, learning and therapeutic criteria (Eysenck & Eysenck 1975, 1985).

The Theoretical Strategy

In the theoretical strategy to structured personality, testing items are selected to assess those variables or constructs that are specified by a major theory of personality. After selecting the items and grouping them into scales, construct-related evidence for validity is sought. In fact, psychometricians have adopted the theoretical strategy in order to avoid the various types of disagreement and biases that usually do stem from factor analytic strategies. The theoretical strategy requires that the items of the test must be consistent with the theory. If the theory states that the personality can be broken down into five major dimensions then the test constructors would strive to write items that tax each of these five dimensions. In this way, this theoretical strategy attempts to create a homogeneous scale and to obtain this aim, they may, use various types of statistical techniques. The following are the major structured personality tests developed using theoretical strategy.

Edwards Personal Preference Schedule (EPPS)

One of the earliest and best-known examples of a theoretically derived structured personality test is the *Edwards Personal Preference Schedule* (EPPS), developed by Edwards (1954, 1959). The theoretical basis for the EPPS is the need theory of personality proposed by Murray and Morgan (1938). In developing the EPPS, Edwards selected 15 needs from Murray's list and constructed items for content validity for each. Each need is paired with other 14 needs twice on the test. There is one additional scale—a measure of response consistency tends to provide a check on random responding. It consists of 15 repeated items. Testees who carefully complete the EPPS generally answer at least 9 or more of 15 items consistently. A score lower than 9 is indicative of potential interpretive problems. The entire inventory consists of 210 pairs of statements in which items from each of 15 needs are paired with items from the other 14. The 15 needs plus one consistency scale have been presented in Table 10.3. Within each pair, the testee is forced to choose one such statement which is more characteristic of oneself. In fact, due to this forced-choice selection, the EPPS results in what is called *ipsative score* which tells that the strength of each need is expressed not in absolute terms, rather in relation to the strength of the testee's other needs. Thus ipsative score presents result in relative terms rather than in absolute terms. Such scores, therefore, compare the individuals against themselves and tend to produce data that reflect the relative strength of each need for that person (Popper 1997). In this way, each person provides his or her own frame of reference.

Since each need is paired twice with other 14 needs, the maximum possible raw score on each scale for an examinee or testee would be $2 \times 14 = 28$. Needless to say, the lowest possible score on a need scale would be zero.

Table 10.3 EPPS Needs and their brief description

n Achievement:	To do one's best and accomplish something difficult
n Deference:	To conform to what is expected of one
n Exhibition:	To be the center of attention, to have others notice oneself
n Intraception:	To analyze the motives and feelings of oneself and others
n Dominance	To influence others and to be regarded as leader
n Nurturance:	To help others in trouble, to treat others kindly
n Order:	To make plans before starting on a difficult task
n Autonomy:	To be able to come and go as desired, to say or do what one wants
n Affiliation:	To be loyal to others, to participate in friendly groups
n Succorance:	To seek encouragement from others, to gain sympathy
n Abasement:	To feel guilty when one does something wrong, to accept the blame when things don't go right
n Change:	To do new and different things, to experience novelty
n Endurance:	To keep at a task until it is finished, to work hard at a task
n Heterosexuality:	To engage in various types of activities with opposite sex, to go out with members of the opposite sex
n Aggression:	To attack contrary points of view, to tell others what one thinks about him
Consistency:	To determine whether or not the responses given on matching pairs are indentical and show consistency

Raw scores on the EPPS are converted into percentile equivalents with the use of various types of norms table representing both college and general adult populations. The strength of the EPPS, particularly in clinical use, is that it measures dimensions of personality that are nonthreatening to the client (Murphy & Davidshofer 1988).

Despite its interesting and impressive features, the EPPS has been criticized and not well received by the reviewers (Heilbrun 1972). Studies have very clearly shown that the effects of social desirability on the test have not been eliminated or even satisfactorily controlled. The EPPS can be easily faked even after use of forced-choice technique. Although the EPPS provides norms for conversion of raw scores into percentiles, the advisability of conversion may be questioned because of the ipsative nature of scores. Ipsative frame of reference is most suitable for *intraindividual comparison* whereas normative reference data are most suitable for *interindividual comparison.* Since the EPPS combines both frames of references, it makes the interpretation of scores more confusing and less meaningful.

Personality Research Form (PRF) and Jackson Personality Inventory (JPI)

Other efforts to use the theoretical strategy in constructing a structured personality test include the development of the Personality Research Form or PRF (Jackson 1967) and the Jackson Personality Inventory or JPI (Jackson 1976a, 1976b). Like the EPPS, the PRF and the JPI were based on Murray's theory of needs.

While developing PRF, Jackson had two major goals. First, he wanted to develop a useful tool for personality research and second, he intended to develop a test for assessing some general personality dimensions assumed to be of importance in environments like schools, colleges guidance clinics, industries, etc. The PRF has been designed basically to fulfil these two basic goals. Initially, the PRF was developed in two sets of parallel forms which differed in length of items and number of scales. The shorter forms (A and B) consisted of 15 scales including one validity scale and 300 items, and the longer forms (AA and BB) consisted of 22 scales including two validity scales (Desirability and Infrequency) and 440 items. The validity scales, like MMPI, were added to detect possible scoring errors, careless test-taking attitudes and social desirability responses. Both sets of parallel forms have been used primarily with college students. This necessitated the development of a newer form (E) of the PRF, which, in fact, represented an extension of the test to other populations. Since it was developed later, it was possible to use a modified item-selection technique that permitted including all 22 scales of the other, longer forms to be converted into a shortened 352 items test. The names of all 22 scales appearing on the PRF are: *abasement, achievement, affiliation, aggression, autonomy, change, congnitive structure, dependence, dominance, endurance, exhibition, harm avoidance, impulsivity, nurturance, order, play, sentience, social recognition, succorance, understanding, desirability* and *infrequency*. Jackson has designed these 22 scales to be bipolar in nature meaning thereby that a low score on any scale may signify not just one absence of trait but also the presence of its opposite. For example, a high score on exhibition scale would mean that the respondent has a passive need to be conspicuous, dramatic and colourful in social situations. A low score on this scale would be indicative of a fear of group activities and active avoidance of such activities.

The Jackson Personality Inventory (JPI) which developed after the PRF through similar, though more refined test construction procedures, has a more appealing and practical orientation (Jackson 1976b, 1994). The JPI has one form which comprises 320 true–false items and 16 scales including one validity scale for use with high school students through college students and adults. The name of scales are: *anxiety, breadth of interest, complexity, conformity, energy level, innovation, interpersonal affect, organization, responsibility, risk-taking, self-esteem, social adroitness, tolerance, social participation, value orthodoxy* and *infrequency*.

Other features of both the PRF and the JPI are that they are balanced in true–false keying and also have no item overlap. Therefore, the scales are relatively independent. On the whole, it can be said that by combining theory-base attempt with rigorous statistical procedures, these tests appear to have established a very scientific trend in the construction of structured personality tests.

The Myers–Briggs Type Indicator (MBTI)

The MBTI has been developed by Myers and Briggs (Myers 1962; Myers & McCaulley 1985). It is a theoretically constructed test based on Carl Jung's theory of Psychological Types (Quenk 2000) and is widely used for assessment of personality in normal individuals. The MBTI is based upon Jung's well-known dichotomy of extroverted and introverted attitudes (E and I) and his classification of opposing ways of perceiving (sensation versus intuition or S vs N), and of contrasting approaches to judgement (thinking vs feeling—T vs F). Besides these, it also includes a polarity of preferences in orientation towards the outer world, (judgement or perception—J or P) which, of course, was not clear in Jung's theory. The scores on these four dimensions, which are presumably independent, results in sixteen possible "type formulas" that present the combinations of the letters of the preferred direction with each of the four dimensions. For example, 'INTP' is one combination type that represents "introverted, intuitive, thinking and perceptive type indicating a person who would be (a) quiet and reserved, (b) enjoy solving problems with logic analysis, (c) interested mainly in ideas, and (d) tend to exhibit sharply defined interests (Myers & McCaulley 1985). Each preferred direction also has a numerical score.

The basic aim of the MBTI is to determine where does an individual fall on extroversion–introversion dimension and on which of the four modes (that is, S, N, T and F) he relies. The underlying assumption of the MBTI is that every person has specific preferences in the way he construes his experiences and these preferences underlie the needs, values, interests and motivation.

Review of studies have demonstrated the applications of MBTI. For example, the MBTI has been used to study the relationship between personality and financial success (Mabon 1998), and sensitivity and different types of purposes in life (Doerries & Ridley 1998). Likewise, this test has been successfully used in the study of leadership (Fitzgerald 1997), emotional perception (Martin et al. 1996), communication styles (Loffredo & Opt 1998) and career choices (McCaulley & Martin 1995).

Measures of Self-concept

Based upon a theoretical strategy, some personality tests have been developed to assess self-concept which broadly refers to the set of assumptions a person has about himself or herself.

Several adjective checklists have been constructed for evaluating self-concept. In these checklists, a list of adjectives is presented and the subjects are asked to indicate those adjectives which apply to them. One such popular adjective checklist is Gough's Adjective Checklist that contains 350 adjectives in alphabetical order from 'absent-minded' to 'zany' (Gough & Heilbrun 1980). The adjectives are based on 15 Murray needs covered by EPPS and two specialized personality theories, namely, Berne's (1961, 1966) Transnational Analysis theory and Welsh's theory of creativity and intelligence. The Student Self-concept Scale (SSCS) is another measure of self-concept based upon Bandura's theory of Self-efficacy (Gresham, Elliott & Evans-Fernandez 1993). The SSCS assesses three major domains of self-concept, that is, academic, social and self-image. Within each domain, the respondents have to indicate not only how confident they are regarding what is being asked by the item but also how important the items are to them and how confident they are that possessing certain attributes or doing something will lead to certain outcomes. The Personal and Academic Self-Concept Inventory (PASCI) is another popular measure of self-concept developed by Fleming and Whalen (1990). It is designed for high school and college students. This inventory attempts to investigate the hierarchical, multifaceted model of self-concept developed by Shavelson and his colleagues. The PASCI, in its current version, consists of a global self-esteem scale and six additional facet scales. Of these six, two deal with social aspects of the self-concept. The Pier–Larris Children's Self-Concept is another measure of self-concept that contains 80 self-assessment items and requires 'Yes' or 'No' response. Besides these, the Tennessee Self-Concept Scale is another measure of self-concept. This is a formal paper-and-pencil test of self-concept.

A relatively novel approach to assessment of self-concept is based on Rogers's theory of self. For evaluating self-concept, Rogers has recommended the use of Q-Sort technique. In this technique, the person has to sort several self-assessments in piles from least to most personally descriptive. He is asked to make two sorts of cards—one type of sort tends to describe who the person really is (real self) and the second type of sort describes what the person believes he or she should be (ideal self). Roger's theory obviously predicts that major discrepancies between these two types of self reflect low self-esteem and poor adjustment (Rogers 1961).

Combination Strategies

In modern times, the test developers are using a combination strategy in which they mix various types of strategies for developing a structured personality test. Two most popular structured tests which rely upon combination strategies are the Million Inventories and the NEO Personality Inventories.

The Million Inventories

Million (1994) has developed a structured inventory designed to assess personality disorders. It is named as the Million Clinical Multiaxial Inventories (MCMI), originally published in 1977. Revised in 1987, it was named MCMI-II and currently again revised in 1994, it has been named MCMI-III. Thus, the Million Clinical Multiaxial Inventories consist of three versions of the same test—the MCMI, MCMI-II and the MCMI-III. Following Piotrowski (1997), the Million Inventories rank second only to MMPI and MMPI-2 as a measure of pathological traits of personality. However, the MCMI-III has two vantage points over the MMPI-2. *First,* it is much shorter (that is, only 175 true-false items) and therefore, more palatable to clinical practices. *Second,* it is well planned and organized to identify the clinical syndromes in a manner that is compatible with the DSM-IV of American Psychiatric Association.

The MCMI was originally published in 1977 and it comprises 24 scales for assessing pathological traits of personality. In fact, the construction of this inventory was deliberately undertaken to meet the criticisms of the MMPI and to maximally utilize intervening advances in the diagnosis of psychopathology. Soon some controversy arose regarding whether the MCMI provided a good measure of personality disorders as defined by the *Diagnostic and Statistical Manual of Mental disorders* (DSM-III). Consequently MCMI was revised and MCMI-II was published (Million 1987). In this edition, Million not only tried to accommodate the personality disorders of the revised versions of DSM-III and DSM-IIIR but also added two new scales to the original 24 scales. Obviously, the basic purpose of the MCMI-II was to help the clinicians make a diagnosis on the basis of Axis II of the DSM-IIIR. Still, controversy continued over how much the MCMI-II really appeared to be an improvement over MCMI. Despite this, it remained in widespread use and following Million's high influential theory of personality disorders which served as one of the conceptual basis in the original formulation of the Axis II personality disorder categories of DSM-III, Some significant changes occurred in this direction.

When the new version of DSM, that is, DSM-IV was published in 1994, Million also revised MCMI-II and produced MCMI-III which comprises 175 brief, self-descriptive statements to be marked 'true' or 'false' by the testee. Of these, 85 items are taken from MCMI-II and 90 items have been taken from the first test (Groth-Marnat 1999). MCMI-III had two additional scales—Depressive personality scale and Post-traumatic stress disorder scale. In fact, MCMI-III is rooted in Million's biopsychosocial views of personality functioning and psychopathology (Million et al. 1996). The score profile of MCMI-III includes 24 clinical scales, apart from four modifying scales (cf. Table 10.4). All these clinical scales are grouped under four major categories, namely, *Clinical Personality Patterns, Severe Personality Pathology, Clinical Syndromes* and *Severe Syndromes.* Clinical Personality Patterns have 11 scales that closely coincide with the Axis II Personality disorders in the DSM-IV; Clinical Syndromes have 7 scales that measures Axis I disorders; Severe Syndromes and Severe Personality Pathology each have three scales. Clinical Personality Patterns and Severe Personality Pathology have scales which are designed to assess Axis II personality pattern disorders of DSM at different levels of severity. Clinical Syndromes and Severe Syndromes have scales which are designed to assess the Axis I syndromes of DSM. In addition to these four major categories, there are modifying indices which consist of four scales (Disclosure, Desirability, Debasement and Validity). Scores on the first three scales are used for adjusting the other score scales in the upward or downward direction, based on defensiveness or exaggeration of symptoms, respectively. The Validity scale assesses the examinee's approach to the test. In fact, this scale consists of three items with unusual content and endorsement to them suggests that the examinee was either not paying attention, responded randomly or was being intentionally oppositional.

Table 10.4 Scales of the Millon clinical multiaxial inventory-III

Clinical Personality Patterns	Clinical Syndromes
Schizoid (1)	Anxiety (A)
Avoidant (2A)	Somatoform (H)
Depressive (2B)	Bipolar: Manic (N)
Dependent (3)	Dysthymia (D)
Histrionic (4)	Alcohol Dependence (B)
Narcissistic (5)	Drug Dependence (T)
Antisocial (6A)	Post-traumatic Stress Disorder (R)
Aggressive (Sadistic) (6B)	**Severe Syndromes**
Compulsive (7)	Thought disorder (SS)
Passive-Aggressive (negativistic) (8A)	Major Depression (CC)
Self-Defeating (8B)	Delusional disorder (PP)
Severe Personality Pathology	**Modifying Indices**
Schizotypal (S)	Disclosure (X)
Borderline (C)	Desirability (Y)
Paranoid (P)	Debasement (Z)
	Validity (V)

Recently, Million has developed two new instruments that extend his approach to the assessment of personality and psychopathology. One is the *Million Adolescent Clinical Inventory* (MACI) and the other is the *Million Index of Personality Styles* (MIPS). MACI was intended to be a tool for clinical use in assessment of adolescents between ages 13 and 19 (T Million, C Million & Davis 1993). MACI has been primarily derived from the Million Adolescent Personality Inventory or MAPI (Million, Green & Meagher 1982) which was an order test designed for use in both clinical assessment as well as vocational counselling and academic advice. MIPS is primarily designed to act as a measure of personality for normal adults who generally seek assistance for family, work or social problems in various counselling situations (Million 1994).

The Million Inventories have reliability and validity similar to other personality tests. Hence, they are psychometrically sound and are in widespread use.

The NEO Personality Inventoy (NEO-PI-R)

The NEO-Personality Inventory-Revised is one of the newest major personality inventories developed by Costa and McCrae (1992). The developers of this inventory have used both factor analysis and theory in item development and test construction. This inventory measures the *Big Five* personality factors: *Neuroticism* (N), *Extraversion* (E), *Openness* (O), *Agreeableness* (A) and *Conscientiousness* (C). 'Big' refers to the fact that each factor subsumes a large number of more specific traits. The Big Five are as broad and abstract in personality hierarchy as Eysenck's

"Superfactors". Originally, Costa and McCrae (1992, 1995) had focused only on three factors like Neuroticism, Extraversion and Openness, thus the title NEO-Personality Inventory. Subsequently, they added the factors of Agreeableness and Conscientiousness to conform to the Big-Five factor model. They differentiated each of these five factors (or domains) into six more specific facets which were defined as more specific traits or components that make up each of the broad big-five factors (cf. Table 10.5). Each facet is assessed by 8 items. Thus the most recent NEO-PI-R consists of a total of 240 items (that is, 5 factors × 6 facets × 8 items). For each item, the testees indicate the extent to which they agree or disagree, using a five-point rating scale.

Table 10.5 NEO-PI-R facet scales associated with Big Five factors

Neuroticism:	anxiety, impulsivity, angry hostility, depression, self-consciousness, vulnerability
Extraversion:	activity, excitement seeking, positive emotions, warmth, gregariousness, assertiveness
Openness to Experience:	actions, ideas, values, fantasy, aesthetics, feeling
Agreeableness:	altruism, trust, straightforwardness, compliance, modesty, tendermindedness
Conscientiousness:	competence, dutifulness, order, achievement-striving, self-discipline, deliberation

The NEO-PI-R has good reliability and validity across different data sources, and with other instruments such as Goldberg's (1992) adjective inventories. Conclusively, it can be said that the NEO-PI-R reflects modern trends in personality test construction by its dependence on logic, theory and liberal use of factor analysis and statistical approaches in test construction. It appears to be exceptionally promising for assessing a wide range of characteristics throughout the different cultures of the world.

On the whole, it may be concluded that psychometricians' attempt to construct various types of structured personality tests is definitely praiseworthy. Of the structured personality tests, the Million Inventories, NEO-PI-R and MMPI-2 are considered to be the dominant tests of the present time.

Self-report Inventories in India

In India, several self-report personality inventories have been constructed. Some of the foreign-made tests have also been adapted to suit Indian conditions. Sohoni (1953) developed a test of temperament and character for high school children. The reliability of the test ranged from 0.44 to 0.54 and the validity coefficient ranged from 0.23 to 0.45. Singh (1967) constructed an adjustment inventory for college students. It measured adjustment in five areas—home, health, society, emotion and education—and had a total of 102 'Yes-No' items. The internal consistency reliability ranged from 0.92 to 0.94. The validity coefficient of the inventory against Asthana's adjustment inventory was 0.62. Bengalee (1964) developed a Multiphasic Personality Inventory which was named the Youth Adjustment Analyser (YAA). The purpose of the inventory was to screen out maladjusted students from the college-going population. It covered five areas of personal and social adjustment, namely, unhealthy parent attitudes, general home adjustment, aggressive behaviour, neuroticism and interests. For measuring the first two personality traits, two independent scales were developed—the Parent Attitude Scale for measuring unhealthy parental attitude and the General Home Adjustment Scale for measuring general home adjustment. The Parent Attitude Scale has five subscales, namely, dominance, acceptance, submission, rejection and total parent attitudes. The General Home Adjustment Scale consists of 38 items and the Parent Attitude Scale consists of 35 items. Items of the former are available in English, Marathi and Hindi languages. Prasad (1974) developed an adjustment inventory for teenagers which measures parent adjustment, home and family adjustment, social adjustment, emotional adjustment and self-acceptance. The inventory has a total of 279 items and norms for different sections of the population. Singh & Sinha (1979) have developed a personality test known as the Differential Personality Scale which measures nine personality traits, namely, decisiveness, responsibility, emotional stability, masculinity, friendliness, heterosexuality, ego-strength, curiosity and dominance. This scale has been revised by Singh & Singh (2002) and has been renamed as Differential Personality Inventory which, apart from the above nine dimensions, also includes the dimension of self-concept. The inventory has a total of 150 items in Hindi and is meant for college students. However, it can be administered in upper classes at school also. The test-retest reliability coefficient for the various dimensions of the scale ranged from 0.73 to 0.86 and the internal consistency coefficients ranged from 0.70 to 0.89. The validity coefficients of the different dimensions ranged from 0.55 to 0.84. The intercorrelations of all the ten dimensions were low and statistically not significant.

There have been some adaptations of foreign-made tests to suit Indian conditions. Mohsin & Hussain (1981) adapted the Bell Adjustment Inventory (Students' form) in Hindi. The Hindi adaptation of the inventory consists of 135 items and in its present modified form published in 1987, has only 124 items. The test-retest reliability coefficients of the four areas of the adjustment of the inventory ranged from 0.700 to 0.926 and the split-half reliability coefficient ranged from 0.738 to 0.932. The validity coefficients of the four areas of adjustment ranged from 0.272 to 0.785 against the Neuroticism Scale of the Hindi adaptation of Eysenck's Personality Inventory, and from –0.088 to –0.255 against the extroversion scale of the same inventory. Bell's Adjustment Inventory has also been adapted in Hindi by Saxena (1959) for the age range of 11–20 years. Singh & Jamuar (1971) adapted the Maslow Security-Insecurity Inventory in Hindi. There are 70 items in the adapted inventory. The test-retest reliability coefficient was 0.79 and the split-half reliability coefficient was 0.86. Percentile norms were developed separately for male and female students of the BA Part-I class. Singh (1972) adapted the Maudsley Personality Inventory in Hindi. The test-retest reliability coefficient for the E scale (Extroversion) was 0.77 and for the N scale (Neuroticism) was 0.82. The inventory was validated against the Guilford–Zimmerman Temperament Survey and several other criteria. The validity coefficients were satisfactory and statistically significant.

OVERCOMING DISTORTIONS IN SELF-REPORT INVENTORIES

Distortions in the actual responses of self-report measures of personality are major problems for users of personality tests. Hence, it is essential that attempts be made to overcome these distortions and make the self-report responses more representative of the true responses. The following measures may be adopted for the purpose:

Establishment of Rapport

In any personality test, distortions occur when the testees feel 'discomfort' and find themselves in an unfriendly environment. It is, therefore, essential that before actual administration of the self-report inventories the tester should make every effort to establish a warm and co-operative relationship called 'rapport' with the testees. The development of such a relationship is, to a large extent, dependent upon the tester's skills and his subtle modification of the testing situation. After a rapport is established, the testees will express the truth in an unhesitant way, thus reducing the major proportion of distortions.

Use of Forced-Choice Technique

The forced-choice technique has been used in controlling faking-good or socially desirable response sets. In forced-choice items, the subject is forced to choose between two or more than two equally desirable or undesirable terms or phrases or statements. The subject who wants to give socially desirable responses is outwitted by the forced choice between equally desirable statements. Several psychologists have used the forced-choice items in controlling faking-good tendency but the results have not been encouraging (Anastasi 1968). Probably, the reason is that forced-choice items have some defects. For example, such items require more time to obtain an equal number of responses. Reliability is also decreased in case of such items because the choice becomes difficult and when reliability is decreased, it may also offset the gain in validity.

Concealing the Main Purpose of the Test

When the subjects do not know the real purpose of the test, it becomes difficult for them to fake, although in such a situation they may be more suspicious and defensive in their responses. Subjects may guess from the nature of items but then, they may not be definite that some inferences about their tendencies or traits are to be made and this will naturally lessen the probability to fake.

There can be two ways through which concealment can be done. One method is to state such a plausible purpose of the test which, in fact, is not the real purpose. For example, a personality test, with utmost care and effort, may be described as a test of ability, and if it appears so to the examinees, the faking is likely to be reduced to a greater extent. Another method of concealment is to insert information which is actually false among items of information which are actually true. For example, the students may be asked to endorse those titles of the book in the booklist which they have gone through. This list of titles will also contain some fictitious titles. The greater the number of endorsements of such fictitious titles, the higher the deceit or boasting.

Theoretically, concealing the real purpose of the test may appear a good means for controlling faking but it may not be considered desirable on ethical grounds because it may produce the impression among people that psychologists are tricksters.

Use of Verification And Correction Keys

Some psychologists have recommended the use of various kinds of correction and verification keys which give an indication whether or not an examinee is presenting a true picture of the self. For example, the famous MMPI uses four validity scores which aim at checking carelessness, evasiveness, misunderstanding and operation of other response sets. Likewise, the Edwards inventory uses several forced-choice pairs of statements of which 15 pairs are presented for the second time at random intervals within the test. A higher degree of inconsistency in choice of responses on the two occasions indicate the operation of some kind of response set and gives a confused picture of the self.

SITUATIONAL TESTS

A situational test represents a sort of compromise between a standardized test and observational methods of assessing personality. As its name implies, personality traits are measured on the basis of observations of ratings of what a person thinks and does in a given situation, which resembles a real situation of everyday life. The person concerned has usually no idea that he is being examined. Ordinarily, the situation represented by such tests is a social situation having opportunity for interaction with other individuals and is specially designed to emphasize those aspects of personality which are under study.

The first situational test of personality was developed by the United States Office of Strategic Services (OSS) during World War II in an effort to screen out men for military assignments.

Situational tests are more suited to the measurement of traits like leadership, dominance, responsibility, extroversion-introversion and the like. Sometimes these tests utilize directly observable units of behaviour as the basis of assessment of the traits. Such situational tests are called *behavioural tests* because they are directly concerned with observable behaviour. Honesty, self-control and co-operation are such traits and all these traits may be grouped under the head—character. One of the first attempts to study these traits was made by Hartshrone, May & Shuttleworth (1930) in their Character Educational Inquiry (CEI). In general, the CEI behavioural tests utilized natural situations lying within the day-to-day routine of a school child such as games, classroom examination, etc. The children placed in these situations were not aware of the fact that they were being studied. The CEI tests were principally designed to measure behavioural traits like altruism, honesty and self-control. However, most of the CEI tests were concerned with measuring honesty among the children by providing opportunity for cheating. These tests utilized different modes or techniques for studying honesty. In one such technique, called the *duplicating technique*, the children were administered one of the common classroom tests like the arithmetical reasoning test or the vocabulary test. A set of duplicate children's responses unknown to them was prepared by the experimenter. In the subsequent administration, the original test was again given with a request to score their own responses with the help of a scoring key. A simple comparison of the responses scored with the duplicated responses revealed whether or not the children had changed their responses in scoring, that is, whether they had cheated or not. Other CEI tests, intended to measure honesty, provided situations in which the person had an opportunity to lie or to steal something.

Likewise, the self-control through CEI tests was measured by the extent to which distraction in doing a task is resisted or the total time the child persisted in solving a given puzzle or a complex task. One such task involved the reading of a story in which the words were all mixed (or run) together. Running over of the words together naturally created reading difficulty and the children's self-control was measured by the length of the time devoted to reading the mixed words. One example of such lines in which words were run together is given below:

THEKINGORDEREDTHEKNIGHTTOBRINGTHECULPRITBEFOREHIM

In fair evaluation to the situational tests, it can be said that although they represent real-life situations and therefore their findings can be easily generalized to natural life situations, they have several important limitations. *First*, situational tests are extremely time-consuming, costly, and laborious techniques. They are time-consuming because ordinarily observations in contrived situations last for several hours; they are costly because they demand the services of trained observers; and they are laborious techniques because creating a real-life situation involves a good deal of labour on the part of observers. *Second*, subjectivity and bias may operate in the observation of the situational tests. Observers, though professionally trained, may not make a fair and objective observation because of certain bias towards the individuals to be observed. Likewise, other subjective elements may enter into his observation. Although an attempt is made to control such subjective elements, they cannot be completely eliminated because the observer is always a human being. *Third*, for increasing the reliability and validity of the observation made under situational tests, the experimenter emphasises on studying an isolated bit of behaviour. The problem arises as to what significance or meaning can be attached to such an isolated bit of behaviour. For example, let us suppose that in a particular situation, 6-year-old Mohan hits 6-year-old Sohan. What is the meaning of this hitting? Does it represent aggression of Mohan towards Sohan? Does it represent an act of dominance of Mohan over Sohan? Or does it represent a sense of superiority of Mohan over Sohan? We are unable to answer if this observation is not studied in relation to other types of behaviours of Mohan and Sohan. Thus, the problem of determining the meaning of an isolated bit of behaviour is always before us in a situational test. *Fourth*, 'What to observe' is also a problem in situational tests, which cannot be planned to study all behaviours of the individual. Only the limited behaviour should be selected for observation. In such a situation, determination and selection of a meaningful set of behaviours to be observed is always an important problem. *Lastly*, whether the observer should be kept visible or invisible is also a problem in most situational tests. If the group is small, the physical presence of the observer in the group may change what actually takes place in the group. How can the observer be fitted into such a setting is always a problem. All these behavioural methods, including the situational tests, convey the full meaning about the traits to be assessed only when they are followed by other standardized tests. Merely on the basis of observational methods, it is difficult to arrive at a particular conclusion. Thus, observational methods may be regarded as only stop-gap arrangements to be replaced by some other more appropriate methods in the future.

MEASUREMENT OF INTERESTS, VALUES AND ATTITUDES

No programme of assessment of personality would be said to be complete and coherent unless it pays due attention to the measurement of interests, values and attitudes of a person.

Meaning and Types of Interest Test

An interest may be defined as a preference for one activity over another. This simple definition reflects two points. *Firstly*, interests involve the selection and ranking of different activities along a like-dislike dimension. One student for example, may select three activities like reading a novel, listening to radio broadcast and watching television and rank them 3, 2 and 1 respectively. Obviously, here he prefers watching a television to listening to a radio broadcast and reading a novel. *Secondly*, interests involve activities or behaviours indulged in by individuals. Interests are often expressed by action verbs as *reading* a novel, *listening* to a radio broadcast, *planting* flowers, *watching* television, etc.

Based upon methods employed to evoke responses, Super & Crites (1962) have classified interests into three types—

(a) Expressed Interest
(b) Manifest Interest
(c) Inventoried Interest

Expressed interest, as its name implies, is the most simple and direct way of obtaining information about interest and is defined as an interest whenever a person states his preference for one activity over another. A teacher may express his interest in reading textbooks over reading novels.

An *interest* is made *manifest* when a person voluntarily participates in an activity. A student who attends Yoga classes clearly demonstrates interest in Yoga and this is his manifest interest.

There is no necessary relationship between expressed interest and manifest interest, though in many situations, they tend to coincide or overlap. Most of the individuals engage in some activities which they claim to dislike and just on the reverse, most people may refuse to engage in the activities which they claim to enjoy.

Inventoried interests are those interests which are measured by interest tests that compare interest in different activities. As we know, most of such inventories measure only a limited sample or set of interests. The interests assessed by such inventories are called inventoried interests.

In the measurement of interests, investigators have concentrated primarily upon vocational pursuits. The most obvious approach to the measurement of interests of a person is simply to directly ascertain his preferences. But experience has shown that the responses given towards such direct questions are often unrealistic and unreliable. This is particularly true when the interest inventory is being administered to young people and children. There are two reasons for this. *First*, young people or adolescents are not fully aware of the different activities that are required in a particular occupation. As a consequence, they are unable to judge whether they would like or dislike a particular activity in a given occupation or area of activity. *Second*, their preferences for occupations are guided by stereotypes which are formed on the basis of the information released by popular mass media like television, radio, magazines, movies, etc. But in reality, things may be quite different. For example, the life of a doctor, a professor or an engineer may be reflected in a particular way to suit the motives of the media. For the reasons stated above, it was realized that a more indirect method of assessing interest should be evolved and as a consequence of this line of thinking, investigations by different psychologists were undertaken. Thus several standardized interest inventories were developed.

In America, the first interest inventory was introduced in 1921 and it was called the *Carnegie Interest Inventory*. Since then, several types of interest inventories have been developed. Some of these interest inventories became famous throughout the world. Here we shall review such famous interest inventories.

The Strong Interest Inventory (SII)

The Strong Interest Inventory was first constructed by E K Strong, Jr., shortly after World War I, that is, around 1920–21. Then it was named as the Strong Vocational Interest Blank (SVIB) which was published in 1927. Subsequently, it was revised several times. In the revised 1966 version of the SVIB, 399 items were related to 54 occupations which were meant for men. A separate form presented 32 different occupations for women.

Beginning in the 1970s, extensive innovations were introduced and implemented in successive revisions of SVIB (Campbell 1974; Campbell & Hansen 1981; Harmon et al. 1994). The following three types of principal changes were introduced:

(a) The introduction of a theoretical framework using Holland's Hexagonal model of personality expressed through vocational choice
(b) The merger of the earlier men's and women's forms and reforming them on new male and female samples
(c) A substantial increase in number of scales for vocational/technical occupations requiring less than a college degree

The 1974 version of SVIB introduced by Campbell is called the Strong–Campbell Interest Inventory (SCII). In its current form, the SCII consists of 325 items and is divided into seven parts (Hansen 2000). The seven parts of SCII are: *Occupations* (131 items), *School subjects* (36 items), *Activities* (81 items), *Amusements* (39 items), *Types of people* (24 items), *Preference between two activities* (30 items), *Your characteristics* (14 items). In the first five sections the examinee records his or her preferences by marking 'Like', 'Indifferent' and 'Dislike'. Items in these five parts fall into categories like occupations, school subjects, activities, leisure activities and day-to-day contact with various types of people. The sixth part requires the examinee to express preferences between paired items (such as dealing with things versus dealing with people). The seventh section asks the examinee to mark a set of self-descriptive statements 'Yes', 'No' or '?'.

There are four levels of scores, differing in breadth. The broadest levels are the six General Occupational Themes scores, the next includes 25 Basic Interest Scales and the most specific level provides the 211 available Occupational Scales. The six General Occupational Themes were based upon Holland's theory of vocational choice. After many years of hectic research, Holland had postulated that interests express personality and that people can be classified into one or more of six categories according to their interests. The General Occupational Themes indentified by Holland's model are named as Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E) and Conventional (C). This is also known as the R-I-A-S-E-C model of Holland. The interest patterns of Holland's six personality factors may be explained in Table 10.6.

Table 10.6 Interest pattern of Holland's six personality factors

Realistic (R theme):	Such persons enjoy outdoor activities and technical materials.
Investigative (I theme):	Such persons are interested in science and the process of detailed investigation.
Artistic (A theme):	Such persons enjoy self-expression and like to be dramatic.
Social (S theme)	Such persons are interested in providing help to others and in activities involving other people.
Enterprising (E theme):	Such persons are interested in power and political strength.
Conventional (C theme):	Such persons have clinical interests and like to be well organized.

Each theme characterizes not only a type of person but also the type of working environment that such a person finds most convenient and congenial.

The second level of scores is for administrative indexes. In fact, these are of less personal importance to the testees and are needed only to ensure that errors were not made in administration and scoring of the test.

The third level of scores provides a summary of a person's basic interests. These scores are designed to provide specific information about the 'likes' and 'dislikes' of the testees. For example, they may suggest whether a person scored high, low or about average in preference for science, and mechanical activities.

The final level of scores is for occupational scales. In fact, these scales occupy an important place in the test profile, which shows the person's score for each of 124 occupations that are broken into six general occupational themes. The scoring for occupational scales are different from the general theme and basic interest scales because the occupational scale compares the testee's or respondent's score with the scores of people working in different professions. The general theme and basic interest scales tend to compare the respondent's score with those of people in general. A recent innovation on SII is the addition of Personal Style Scales (Harmon et al. 1994). These scales basically aim to assess preferences for broad styles of living and working. The four personal style scales are:

(i) **Work Style Scale:** High score on this scale indicates a preference to work with people and low score indicates an interest in ideas, data and things.

(ii) **Learning Environment Scale:** High score indicates a preference for academic learning environments and a low score indicates preference for applied learning activities.

(iii) **Leadership style scale:** High score indicates comfort in taking charge of others and low score indicates uneasiness in such activities.

(iv) **Risk taking/Adventure scale:** High score indicates a preference for risky and adventurous activities and low score indicates preference for safe and predictable activities.

All these four personal style scales help in vocational guidance by displaying level of comfort with distinctive styles.

Researches have provided sufficient evidence to the fact that the interests measured by the various tests are stable (Kaplan & Saccuzzo 2001). All scores on the Strong Inventory are described in terms of standard scores with a mean of 50 and standard deviation of 10. The reliabilities as well as concurrent and construct validity of the test are highly satisfactory.

The Kuder Occupational Interest Survey (KOIS)

The Kuder Occupational Interest Survey (KOIS) ranks second in popularity and tries to provide a unique alternative to SCII. The earliest version of KOIS was named the Kuder Preference Record-Vocation in which forced-choice triad items were used and the respondents indicated which of the three activities they would like most and which least. In this original version the scores were obtained not for specific occupations but for 10 broad interest areas, namely, Outdoor, Mechanical, Computational, Scientific, Persuasive, Artistic, Literary, Musical, Social Science, and Clerical. Subsequently, this inventory was revised and a downward extension of the Kuder Preference Record-Vocation was made. This revised inventory was named the Kuder General Interest Survey (KGIS) which was designed for Grades 6 to 12, using simpler language and easier vocabulary. A still later version of KGIS appeared and was called the Kuder Occupational Interest Survey (KOIS) and it provided scores with reference to specific occupational groups as does the Strong Inventory (Kuder & Diamond 1979; Kuder & Zytowski 1991). Unlike Strong, on the KOIS, the respondent's score on each occupational scale is expressed as a correlation between his/her interest pattern and the interest pattern of the particular occupational group. The KOIS now provides both occupational scores and 10 broad homogeneous basic interest scores called *Vocational Interest Estimates* (VIE) which are, in fact, percentile scores derived from short scales equivalent to 10 interest area scores of the early Kuder Preference Record. They can also be converted into R-I-A-S-E-C theme of Holland by establishing direct correspondence for some scales and by averaging percentiles on two or three Kuder scales for others (for example, average of Kuder Artistic, Literary, and Musical for Holland's Artistic theme).

Studies show that the psychometric properties of KOIS are very satisfactory. Short-term reliabilities tend to be high (between .89 and .95) and also, evidence suggests that scores remain stable for as long as 30 years (Zytowski 1996). Reviewers have however, pointed out the dearth of the predictive validity data of this scale. Apart from this, the inventory also fails to address the effects of its forced-choice format on scores (Herr 1989; Tenopyr 1989).

Jackson Vocational Interest Survey (JVIS)

The Jackson Vocational Interest Survey was developed by Jackson (1977). It consists of forced-choice type items where the respondents must indicate their preference between two equally popular interests. Rational classification of vocational interest items were of two types: One set was defined in terms of *work roles* and the other set was defined in terms of *work styles*. Work roles relate to what a person does on the job and work styles refer not to job-related activities but to preferences for particular working environment or situations in which a certain kind of behaviour is expected. Examples of work styles include independence, playfulness, etc., which were, of course, related to a person's values either directly or indirectly.

JVIS has been revised in 1995 and the final form contains 34 basic interest scales, covering 26 work roles and 8 work styles. It is designed in such a way that it could be equally applicable to both sexes, although separate percentile norms for male and female are available.

Today JVIS is basically used for the career education and counselling of high school and college students. It is also used in planning careers for adults. It contains 289 statements describing various job-related activities. It takes about 45 minutes to complete and from 34 basic interest scales, scores in 10 General Occupational Themes are derived: Expressive, Logical, Inquiring, Practical, Assertive, Socialized, Helping, Conventional, Enterprising, and Communicative.

Psychometric properties of JVIS are sound. Reliability for ten occupational themes is about .89 and the test-retest reliability of 34 basic interest scales ranges from .84 to .88. Validity studies have suggested that JVIS predicts the interest pattern of university students better than any other measure. JVIS has both hand-score forms as well as machine-score forms.

The Career Assessment Inventory (CAI)

The Career Assessment Inventory (CAI) had been first developed by Johansson (1976) and currently revised in 1985. It is based on the Strong Inventory. It is designed specially for persons who seek a career but don't have college degree or any advanced professional training. It basically concentrates on skilled trades, technical work and various types of service occupations.

The CAI has 305 items which are grouped into three content categories: *Activities, School Subjects* and *Occupations*. In each item, there are five response options ranging from 'like very much' to 'dislike very such'. Although CAI is written at a sixth-grade reading level, it can also be used with adults who have poor reading skills. It provides scores on three major types of scales, including Holland's six General Theme scales, 22 homogeneous Basic Interest Areas scales and 89 occupational scales (Kaplan & Saccuzzo 2001). Administrative Indices and four nonoccupational scales have also been included.

Self-Directed Search (SDS)

The Self-Directed Search (SDS) has been developed by Holland (1985, 1992) whose hexagonal model of general occupational themes has attracted wide attention. It is a unique interest inventory in the sense that it is a self-administered, self-scored and self-interpreted, vocational counselling instrument (Spokane & Catalano 2000). The SDS contains 228 items. A set of 66 items, grouped into six scales with 11 items each, describe activities. A set of another 66 items assess competencies which are again grouped into six scales of 11 items each. Occupations are evaluated by six scales of 14 items each. Self-estimates are obtained in two sets of six ratings. The respondent fills out the SDS, scores the responses and calculates six summary scores, corresponding to the themes of the Holland model. These six summary scores can be used to obtain codes that reflect the highest areas of interest. The SDS also attempts to simulate the process of counselling by allowing respondents to allow to list occupational aspirations, then indicate occupational preferences in six areas and ultimately, rate abilities and skills in these areas. Using the SDS, the respondents can also build a meaningful *personal career theory* which, going beyond interests, includes readiness for career-decision making and readiness to obtain guidance (Reardon & Lenz 1999). Recently, the SDS has also acted as a link to an occupationl finder. In the 1994 edition of the SDS, the respondent can locate over 1300 occupations and match his interest code to corresponding occupational choices.

Of these models or interests, Kuder's model seems more suitable for Indian conditions. Singh (1965) adapted the Kuder Preference Record (Vocational) in Hindi; Naik (1969) adapted the same form in Oriya; Parikh (1971) adapted the vocational form in Gujarati; and Gopalan (1972) adapted it in Malayalam. Trivedi (1969) constructed an interest inventory for undergraduate students. Palsane (1975) also constructed an interest inventory in Marathi for school and college-leaving students.

Value Tests

Value tests intend to measure generalized and dominant interests and in this way, they are different from interest tests, which intend to measure interest in specific activities or areas. Value tests are mainly concerned with preferences for 'different ways of life'. In general, values refer to a shared, enduring belief about ideal modes of behaviour or end states of existence. Values shape attitudes, instill actions and determine efforts to influence others. Values also arise in response to the societal and cultural conditions. Psychologists have paid comparatively little attention towards the development of value tests as compared to interest tests. As a consequence, no satisfactory instruments for assessing human values are available. However, this does not mean that no tests of values have been constructed so far. A very popular measure of values is the *Study of Values* (SOV) developed by Allport, Vernon & Lindzey. The study of values is based upon the six categories of values, which were originally suggested by Spranger's *Types of Men* (1928).

The six categories of values are: theoretical (T), economic (E), aesthetic (A), social (S), political (P) and religious (R). According to Spranger the generalized and dominant interest of the theoretical person is *the discovery of truth*; the generalized and dominant interest of the economic man is in *what is useful and practical*; the aesthetic man is interested in giving the highest value to *form and harmony and to the enjoyment of each unique experience gracefully*; the dominant interest of a social person is *love of people*; the political person places the highest value on *power and influence* and the religious person places the highest value on *the unity of all experiences in a bid to understand the cosmos as a whole*. Items in the study of values present problem situations, under each of which the person is asked to choose either from two alternatives or from multiple alternative responses. The responses become the index of the relative strength of the six basic values.

Readers should not assume that a person can be wholly placed under any one of the above categories. In reality, a person may be classified under more than one category and subsequently, his most dominant interest (that is value) may be determined. Studies have revealed sex differences: Male adults tended to shift toward a T-E-P profile whereas female adults tended to shift toward an A-S-R profile.

Rokeach (1973), accepting a debt to Allport and the study of values, developed another important value test called *Rokeach Value Survey* (RVS). This test incorporates two broad kinds of values: *Instrumental values* and *Terminal values*. Instrumental values refer to the derivable modes of conduct whereas terminal values refer to the desirable end states of existence. Ambition is an example of an instrumental value and family security is an example of terminal value. The RVS assesses 18 instrumental values and 18 terminal values. All these 36 values are listed in Table 10.7.

Table 10.7 The 36 values constructs of Rokeach value Survey (RVS)

	Terminal values		
1.	A comfortable life	10.	Inner Harmony
2.	An Exciting life	11.	Mature love
3.	A sense of Accomplishment	12.	National Security
4.	A World at Peace	13.	Pleasure
5.	A World of Beauty	14.	Salvation
6.	Equality	15.	Self-respect
7.	Family Security	16.	Social Recognition
8.	Freedom	17.	True Friendship
9.	Happiness	18.	Wisdom
	Instrumental values		
1.	Ambitious	10.	Imaginative
2.	Broadminded	11.	Independent
3.	Capable	12.	Intellectual
4.	Cheerful	13.	Logical
5.	Clean	14.	Loving
6.	Courageous	15.	Obedient
7.	Forgiving	16.	Polite
8.	Helpful	17.	Responsible
9.	Honest	18.	Self-controlled

The examinees are asked to rank separately the 18-terminal and 18-instrumental values based on their importance for self as the guiding principles in life. The rank for each item becomes the score for that value. Ties are not allowed and therefore, value scores will range from 1 to 18. The lower scores indicate greater importance.

The RVS has some limitations. *First,* the survey possesses marginal reliability which automatically means that the test should not be used for individual guidance. *Second,* the survey ignores several other important values like physical well-being, individual rights, thriftiness and carefreeness. *Third,* individual values are not defined in detail and Rokeach has shown excessive dependence upon a single item for each value instead of using multi-item indices for the different value constructs.

Some value tests have also been developed in the Indian condition. Prasad (1968) has developed an inventory of vocational values for undergraduate and the postgraduate students. The inventory covers nine areas such as altruism, physical conditions of the work, power, prestige, security, economic returns, self-enhancement, social climate and tradition. The inventory has a total of 72 pair items. The test-retest reliability coefficient of the inventory ranged from 0.80 to 0.92 and the internal consistency coefficients ranged from 0.90 to 0.92. Bhatnagar & Tandon (1975) and Verma (1972) have adapted the Allport–Vernon–Lindzey Study of Values in Hindi. Nazre (1968) adapted the Survey of Interpersonal Values (SIV) in Hindi. Value tests based upon the Spranger Type have also been developed in Hindi by Singh & Sharma (1975) and by Ojha (1975). Chauhan & Tiwari (1975) also developed a value test (dimensional), which measures cosmopolitanism vs localism, democraticism vs aristocratism, scientism vs fatalism, venturesomeness vs nonventuresomeness, etc.

Measures of Attitudes

An attitude is best defined as an enduring system of the cognitive component, the feeling component and the action tendency component, all of which centre round an object, person, event, etc. The cognitive component relates to belief regarding the object of attitude, the feeling component relates to the emotions regarding the object of attitude, and the action tendency component (also known as the behavioural component) relates to the action or behavioural readiness associated with the object of attitude. It is obvious from the above interpretation that an attitude has a well-defined object of reference. The degree of a person's attitude may vary from favourable through neutral to unfavourable.

There are numerous measures of attitudes. One of the most popular measures of attitudes is the self-report inventory where there are a large number of statements, which directly ask the persons what their attitudes are. Such self-report inventories are known as *attitude scales.* Thus attitude scales consist of a set of statements to be responded to. It is common for a person to hesitate when he is asked to express his attitude directly. Not only this, sometimes he gives it a social colour, too. In order to remove these difficulties, some indirect measures of attitude have also been developed.

In the measurement of attitudes, the following two underlying assumptions are basic:

1. In any study of attitude measurement, it is assumed that an individual's behaviour with respect to the object of attitude will be consistent from one situation to another. If, for example, an individual dislikes co-education, he would continue to dislike co-education, in all situations. If consistency is not exhibited, it is difficult to assess the individual's attitude.

2. Attitude cannot be observed directly. It is, therefore, assumed that it must be inferred from the statements and actions of a person, that is, it must be inferred from behaviour—verbal or nonverbal—of the person. The statements and actions towards the object of the attitude reveal the degree of favourableness or unfavourableness, which together constitute the valence of the attitude.

As mentioned earlier, there are numerous direct and indirect measures of attitude. Campbell (1950) has categorized them under four major headings.

Nondisguised Structured Tests

These tests attempt to measure attitude on the basis of direct statements regarding the object of attitude. The Thurstone Scale, Likert Scale, Guttman Scale, etc., are a few examples of this category. (A detailed discussion of these various attitude scales is presented in Chapter 13.)

Nondisguised Nonstructured Tests

Under this heading, the following methods of assessing attitude are included:

1. *Survey interview*: Survey interview is another nondisguised nonstructured technique to measure the attitude of a person. Survey interview aims at interviewing as many persons as possible in a short time. Interviewers usually put any of two types of questions—fixed-alternative questions and open-ended questions. In the fixed-alternative questions, the person must choose his answer from the given alternatives. No provision for other categories of answer is made. Two examples illustrating the fixed-alternative questions are given below.

(a) Which party do you think would do a better job of improving the law and order situation in the country in the next few years? (i) Congress (I) (ii) BJP (iii) RJD

(b) Which party do you think has the largest number of honest leaders? (i) Congress (I) (ii) BJP (iii) CPI

The open-ended questions have no prescribed alternatives. Here persons are free to respond in whatever terms and frames of reference they like. The following examples illustrate the open-ended questions:

1. Why did the Janata Party suffer a defeat in the 1980 mid-term Lok Sabha poll?
2. What do you think about the prospects of the Congress (I)?

In the open-ended questions, the interviewer usually adopts nondirect means for getting full answers and co-operation from persons. He encourages an individual to talk freely and, if necessary, uses some mild probes like "Would you like to say more?" or "Why?" or "How?", etc. Both the fixed-alternative and the open-ended questions have been frequently used in assessing attitudes. Since fixed-alternative questions have certain advantages like simplicity and scoring economy, they usually yield more reliable data than the open-ended questions. But this does not mean that the open-ended questions should be abandoned in favour of the fixed-alternative questions. In fact, both types of questions can be used together—the open-ended questions for exploring general problems and contexts of the object of attitude, and the fixed-alternative questions for exploring specific problems and contexts of the object of attitude.

2. *The biographical and essay methods*: In this method attitude is assessed from the opinions expressed in letters, autobiographies, diaries, written or oral interviews, etc. In written interviews, sometimes the persons are asked to write short essays concerning the objects of attitude. Opinions, revealed through all these sources, help the investigator in assessing the attitude of the persons. One general demerit of these methods is that while writing letters or an autobiography or biography, the persons sometimes do not express the true content of a humiliating event because they know that such materials may further expose them to humiliating situations.

Disguised Nonstructured Tests

Under this heading the typical projective tests, originally developed to measure personality variables like want, need, anxiety, etc., are employed. A few examples of the projective tests are the Rorschach test, TAT tests, the Word-association test, the World test, etc. (A detailed discussion of projective tests is presented in Chapter 11.) One of the basic assumptions underlying these projective tests is that when the individual makes a specific interpretation about the situation presented through these techniques, the interpretation is not determined by something in the picture itself but by something about himself (his motives, attitudes, values, etc.). Proshansky (1942) devised one such projective test for studying attitudes towards labour. From various sources, he collected some pictures which obviously had something to do with working people in conflicting situations. Subjects were shown these pictures with the instruction to describe them in as much detail as possible. They were also asked to give their details in the form of a story. Subsequently, three qualified judges rated each story as to the degree of favourableness or unfavourableness of attitude towards labour. Proshanky found a very high agreement between the attitude evaluated by the judges and the picture responses. He, however, found two serious limitations in projective-type tests. *First,* the descriptions of the pictures given by the individuals are sometimes so vague and ambiguous that scoring becomes difficult. *Second,* judges are influenced by their personal biases in ratings. These two limitations make the projective-type tests of attitude unreliable and invalid.

Disguised Structured Tests

Under this heading again, the indirect methods of assessing attitudes are included. These indirect measures are different from the projective tests and very closely resemble the objective testing of attitudes as it is found in the self-report inventories. One example of such a disguised-structured test is the error-choice technique developed by Hammond (1948) to measure attitudes towards labour management and towards Russia. The technique requires a number of nonfactual items (or error-choice items), each of which allows two alternatives, out of which one that seems to be the correct one, is to be selected by the subject. In a test of attitudes towards labour, one of the items was like this:

The average weekly wage of the war worker in 1945 was...... . (a) $37, (b) $57.

In reality both the above averages were wrong but as the subject was forced to choose one, and choosing from the alternatives depended largely upon the subject's attitude. Likewise, in a test of attitude towards Russia, one of the items was as under:

Russia's removal of heavy industry from Austria was...... . (a) legal, (b) illegal.

In the above item or similar other items, facts were really indeterminable and hence a subject's response of choosing either "legal" or "illegal" revealed his own attitude towards Russia. Hammond prepared a large number of nonfactual items (or error-choice items) and subsequently, presented them to the subjects. The labour test and the Russia test were administered under the guise of an 'information test', that is, the subjects were told that the test was merely intended to gather information and therefore, their co-operation was needed. The systematic errors in checking the alternatives on error-choice items in one or the other direction were taken as a measure of attitude towards the object. Hammond had recommended that for effective use of the error-choice techniques, the number of alternatives might be increased from two to three or five. Kubany (1953) used the error-choice technique for measuring attitudes towards national health insurance and found that the indirect technique made more valid distinctions among persons holding favourable or unfavourable attitudes towards national health insurance.

It is obvious from the above discussion that the first two measures (nondisguised structured tests and nondisguised nonstructured tests) are direct measures of attitude, whereas the last two measures (disguised nonstructured tests and disguised structured tests) are indirect measures of attitudes. The indirect measures of attitudes have two principal advantages over the direct measures. *First,* indirect measures enable the investigator to measure the attitude of the individual without his being aware. It has been observed that when individuals are aware that their attitudes are being assessed (as is the case with direct measures of attitude), this feeling in itself produces a change in the attitude being measured. Thus, indirect measurements enable the investigator to measure an attitude without any effect upon the attitude itself. *Second,* for assessing attitudes towards an object or event, which is considered more or less confidential or private and which a person may not want to reveal in public, indirect measures are more reliable and valid. For example, when the investigator is studying the attitude of unmarried mothers towards sex, indirect measures would yield more valid data than direct measures because such mothers violate social norms and are considered to be stigmas in Indian society.

☑ Review Questions

1. Discuss the research importance of self-report inventory and situational tests as tools of personality assessment.
2. Write a review of some of the important self-report inventories developed in Indian conditions.
3. Throw light upon the importance of interest tests and value tests in psychological researches.
4. Discuss the research utilities of MMPI and Bell Adjustment Inventory in psychological researches.
5. Which measures of attitude do you consider best? Give reasons.

◆

11

PROJECTIVE TECHNIQUES

CHAPTER PREVIEW

- Meaning and Types of Projective Techniques
- Classification of Projective Techniques
- Pictorial Techniques
- The Rorschach Test
- Interpretation of the Rorschach Protocol
- The Holtzman Inkblot Test
- Thematic Apperception Test, or TAT
    - Derivatives of TAT
- Verbal Techniques
    - Word-Association Test
    - Sentence-Completion Test
- Expressive Techniques
    - Figure-Drawing Tests
    - Toy Tests
    - Artistic Productions
    - Graphology
- Evaluation of Projective Techniques
    - Fakability
    - Objectivity
    - Standardization
    - Reliability
    - Validity
    - Situational Variables
    - General Applicability

MEANING AND TYPES OF PROJECTIVE TECHNIQUES

Projective techniques which originated in a clinical setting are the indirect measures of personality. The history of projective assessment goes back to the 1400s when Leonardo da Vinci is said to have selected pupils on the basis of their attempt to find shapes and patterns in ambiguous form (Piotrowski 1972). In modern time Galton developed the first projective technique, a word association test, in 1879. This procedure was adapted by Kent and Rosanoff into testing in 1910 and it was frequently used by Jung in therapy. These informal projective techniques gradually evolved into projective tests. In fact, Frank (1939) had introduced the term projective method for describing a category of tests for studying personality with unstructured stimuli. In projective tests, the individual is given an unstructured situation to which he responds. By an unstructured situation, we mean a situation whose meaning and interpretation vary from individual to individual. Such situations have no right or wrong answers and are capable of evoking fantasy material from the testees (Lindzey 1961). The most important assumption of projective techniques is that while responding to an unstructured situation, an individual projects his own feelings, needs, emotions, motives, etc., (which are mostly latent and unconscious) without being aware of doing so. Since the individual is not aware of these revelations, he doesn't resort to any defensive reactions. Thus, in a projective test, the individual has ample opportunity to project his own personality attributes that are mostly latent and unconscious in the interpretation of an unstructured situation. Such latent and concealed experiences are generally incapable of exposure by the questionnaire type of test.

Projective tests differ from the self-report inventories and observational method of assessing personality. Self-report inventories provide a direct and structured situation for assessing the traits of personality, whereas the projective tests provide an indirect and unstructured situation for assessing these traits. In self-report inventories attention is focused upon the measurement of traits, that is, upon the measurement of personality traits individually, whereas projective tests are characterized by the global approach to the assessment of personality, that is, in the projective tests attention is focused upon the measurement of personality as a whole. Projective tests are also different from the observational methods, which are primarily dependent upon the fact that the observer must know the person being observed in detail and he must report the observation without any bias. Projective tests require no such observer and the individual, to a greater extent, gives an objective report of his observation of an unstructured situation.

CLASSIFICATION OF PROJECTIVE TECHNIQUES

There are different types of projective techniques, and different psychologists have classified them into different categories. The earliest classification of projective techniques was done by Frank (1939). His classification is mainly based upon the nature of response evoked by the materials of projective techniques.

On this basis, he classified projective techniques into five categories as given below.

Constitutive

This category includes all those test situations in which the examinee constitutes or frames structures upon materials, which are yet unstructured. Finger-painting and drawing-completion are its best examples. According to Zubin, Eron and Schumer (1965) the Rorschach test also falls in this category in the sense that here the examinee imposes his own structure upon unstructured inkblot situations.

Constructive

The constructive category, though apparently similar to the constitutive category, includes all those test situations where the examinee is required to construct a specified task. He is required to impose a degree of structure upon the situation in the direction specified by the examiner. The examinee may be asked to draw a figure of a human male or female and this would be included in the constructive category rather than in the constitutive category where drawing according to the examinee's own wish or desire is done. Thus, the constitutive category test situations allow a free expression of the examiner's inclination whereas the constructive category test situations do not give such permission.

Interpretative

This category includes all those test situations where the examinee is required to add a comprehensive meaning to the situation. The Thematic Apperception Test (or TAT) and the word association test are included under this category.

Refractive

Under this category are included all those techniques through which the examinee is given an opportunity to express his personality in the form of painting, drawing, handwriting, etc. Graphology (or handwriting) has been cited by Frank as the best example of this category.

Cathartic

This category includes those situations whereby the examinee is given an opportunity through some manipulative tasks for the release of his conflict, wish, etc. Play techniques are cited as its best example.

In fair evaluation of Frank's classification, it can be said that his classification solves less and creates more problems. One of the biggest limitations is that his classification overlaps so much so that the same test can be included under more than *one or two* categories. For example, a figure-drawing test may be included in either the constitutive or constructive or refractive category. Frank's classification, therefore, is not a widely accepted classification of the projective test.

The more convincing classification of projective techniques has been recently provided by Lindzey (1959). Based upon the responses of the examinees, he has divided projective techniques into the following five categories:

Association Techniques

This category includes all those situations where the examinee is required to respond with the associations which are evoked in his mind after seeing or listening to stimulus materials. The Rorschach test, the Holtzman Inkblot test and the word-association test are its best examples. The Rorschach test requires the examinee to respond to an unstructured situation of inkblots in the form of verbal associations with objects, events, persons, etc. No attempt is made to mould those associations either by the examinee or by the examiner. Similarly, in the word-association test the examinee is presented with a variety of words one after another with the instruction to respond with the very first word that comes to his mind after listening to the stimulus word. Subsequently, the responses and the reaction time (the time elapsing between the presentation of the stimulus word and the response word) are analyzed for studying the personality.

Construction Techniques

This category includes all those situations where the examinee is required to construct a story after seeing the stimulus materials (usually a picture) within a certain specified time. No record is generally kept of time but the examinee's themes and mode of responding are considered relevant. The Thematic Apperception Test (or the TAT), the Children's Apperception Test (or the CAT), the Blacky Pictures, the Object Relations Technique and the Pickford Projective Picture (PPP) are some of the best examples of construction techniques. In all these tests, the examinee is required to construct or produce simple or complex statements in the form of a story. A discussion of all these tests is beyond the scope of this book. However, some of them appear in detail in later sections of this chapter.

Completion Techniques

These techniques include those situations where the examinee is presented with some incomplete sentences with the instruction to complete them in any way he desires. A few examples which illustrate the techniques are given below:

I feel tense ....
My ambition in life is ....
I often get nervous ....

Responses given by the examinee are interpreted and analyzed to find some clue regarding his personality. These techniques, however, lack a uniform and standard mode of analysis. Stories have also been used as completion techniques. The Madeleine Thomas Completion Stories test (Mills 1953) and Rotter's Sentence Completion test are examples. The Rosenzweig Picture-Frustration Study (Rosenzweig 1949) is another example of completion technique.

Expressive Techniques

This technique includes those situations where the examinee expresses his personality through some manipulative tasks, which usually involve some interaction with given materials. Play, drawing, role-playing, painting, finger-painting, etc., are the common expressive techniques. One important feature of expressive techniques is that the examiner pays much attention to the way or process by which the examinee manipulates the given materials. For examples, he may ask the examinee to play with a given set of dolls; he may pay attention to the process by which the dolls are selected and handled during the play. Thus in the expressive techniques, attention is given to the process and not to the end product of the process. Expressive techniques, in this way, are different from construction techniques because here much emphasis is given to the process or way of handling the test materials rather than upon the end product of the process (such as the content or theme of the stories, etc.). The famous Toy-World Test is an example of expressive technique.

Choice Techniques

Choice techniques (also known as Ordering techniques) are not projective techniques in the true sense of the term; rather they may be regarded as a step towards objectifying the projective techniques (Kerlinger 1973). Usually, the examinee is presented with some sets of pictures or items (which convey the different degrees of a trait) with the instruction to choose the most relevant and appropriate picture. Sometimes, he may be asked to order or rank those pictures in terms of his preferences and hence the name *ordering techniques*. The choice of the subject becomes the basis for the inference regarding his personality. The Szondi test is an example of ordering technique because the examinees are required to rank sets of pictures along a like-dislike dimension.

Still another important classification of the projective techniques has been done by Best (1978). He has proposed a four-way classification of projective techniques as stated below:

Association

Under this category are included all those techniques where the examinee is presented with stimulus materials like picture, inkblots, words, cartoons, etc., and is asked to tell what he sees, thinks and feels. The Rorschach test, the Holtzman Inkblot test and the word-association test belong to this category.

Completion

This category incorporates the situation where some incomplete sentences are presented to the examinee who completes them in the way he likes. The sentence-completion test belongs to this category. Such tests have been found to be most useful in interpreting traits like anxieties, guilt-feelings, hostility, aggression, attitudes toward sexuality, paranoid tendencies, etc. Items illustrating paranoid tendencies may be framed like this:

Somebody is always trying ....
I worry about ....
I think that my body ....

Role-playing

This technique requires the examinee to act out a specific role in a group of two or more individuals for a certain period of time. It has been commonly observed that when an examinee is asked to play a certain role, he says surprising things, which he may never have said otherwise. His sayings reveal the underlying motives, wishes, needs, attitudes, etc. The role-playing technique has been found to be most suitable for traits like hostility, sympathy, frustration, dominance and authoritarianism.

Creative or Constructive

This technique includes all those test situations where the examinee is required to do some tasks like playing with dolls or toys, paint or finger-paint and draw a figure, etc. Best also says that sometimes the examinee may be asked to write imaginative stories about some assigned situations. Hence, TAT is also included under this head. Generally, on the basis of the choice of colours, words, forms, order of the sentences, etc., an inference regarding the traits of the personality of the examinee is drawn.

A comparative study of Lindzey's classification and Best's classification makes it clear that the former is more appropriate than the latter. The role-playing and the constructive techniques of Best are covered under one category, that is, the expressive techniques of Lindzey. Besides this, Lindzey's classification is very broad and fact-covering. Almost all the projective techniques are covered under any one of the categories without much overlapping.

One way to reduce the confusion and overlapping in classification of projective techniques is to classify them into three broad categories: *the pictorial techniques*, the *verbal techniques* and the *expressive techniques*.

Pictorial

These techniques include all those situations where the unstructured situation consists of vague and ambiguous pictures and the examinee is to respond towards those pictures. His response may be in terms of a few words as it is done in the Rorschach test and Holtzman Inkblot test, or in terms of a series of sentences as it is done in the TAT, the CAT, the Rosenzweig Picture Frustration test, etc.

Verbal

In this technique, the stimulus materials are presented in verbal form such as in terms of words or incomplete sentences. The examinee may be required to respond to those words with the first word that comes to his mind and in case of incomplete sentences, he is required to complete them in a way he likes. Thus the verbal techniques are different from the pictorial techniques in the sense that here the stimulus materials are verbal and not pictorial. The word-association test and the sentence-completion test are its best examples.

Expressive

This technique incorporates all those situations whereby the examinee is given an opportunity to express his personality mostly through some manipulation and objective tasks. Graphology, painting, drawing, playing with dolls and role-playing are some of the common expressive techniques. The *way* or the *manner* in which the examinee does his task in the above-mentioned situations becomes the basis for appraisal of his personality traits.

Thus, we see that there are different types of projective tests. These varied projective tests have the following characteristics in common:

(i) All projective tests share one common assumption called projective hypothesis, which states that a person will project something important about himself or herself on to a vague or ambiguous stimulus usually called unstructured stimuli.

(ii) The procedures used in the projective tests are usually disguised to some degree. In other words, the testees do not know the way their responses will be analyzed and interpreted.

(iii) Projective tests are relatively *unstructured* and therefore, the task incorporated generate inexhaustible variety of possible responses.

(iv) In projective tests, the instructions given to the testee or examinee are general and are often vague themselves. Instructions like 'Tell me what you see' or 'There is no right or wrong answer', or 'what might this be?' are common.

(v) Projective tests usually employ a *global approach* for assessing the person's entire personality. In other words, the psychologist puts together a comprehensive description of the testee's or the examinee's psychological functioning rather than only identifying a particular trait.

(vi) In clinical situation, the projective tests are used in an *idiographic manner*, that is, projective tests in such a situation are used to develop a model of personality functioning of an individual.

(vii) Supporters of the projective tests state that these techniques are only effective means for evaluating unconscious elements of personality. Here, the basic assumption is that we get such information through projective tests that cannot be gathered otherwise.

(viii) The interpretation of responses of projective tests has been strongly influenced by psychoanalytic thinking. The concept of psychic determinism and the importance placed upon the various unconscious processes paved the way for projective testing.

Thus despite its varied types, the projective tests have some degree of commonality.

PICTORIAL TECHNIQUES

As mentioned earlier, pictorial techniques are those techniques where the unstructured situation consists of vague and ambiguous pictures to which the examinee has to give a response. The responses may be in terms of words or a series of sentences. The Rorschach test, the Thematic Apperception test and The Holtzman Inkblot test are the best examples of pictorial techniques. A detailed discussion of these are given below.

THE RORSCHACH TEST

The most popular projective technique is the Rorschach Inkblot test developed by Hermann Rorschach, a Swiss psychiatrist in 1921, to make a diagnostic investigation of personality as a whole. At the outset, it must be borne in mind that the the Rorschach test is a measure of both the *intellectual* and *nonintellectual* traits of personality. Rorschach investigated with a large number of inkblots out of which only 10 inkblots that differentiated most between various psychiatric syndromes were selected to constitute a test. Thus the Rorschach test consisted of ten cards, each of which contained a bilaterally symmetrical printed inkblot. Five inkblot cards (Cards I, IV, V, VI, and VII) are made in shades of black and gray; two cards (Cards II and III) contain bright patches of red in addition to the shades of black and gray and the remaining three cards (Cards VIII, IX and X) contain several pastel shades. An inkblot like that used in the Rorschach test is illustrated in Figure 11.1. Major life events of Hermann Rorschach have been given on the next page.

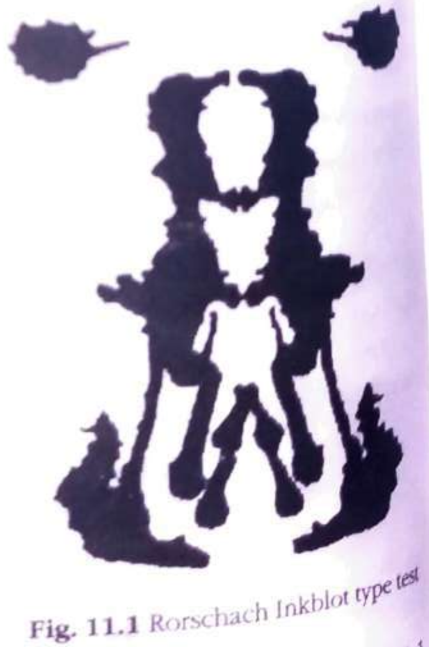

Fig. 11.1 Rorschach Inkblot type test

Major life events of H Rorschach

1. H Rorschach was born on 4 November 1884 at Zurich, Switzerland.
2. Rorschach visited Russia in 1906 and started loving Russia and Russians.
3. In 1910, Rorschach married a Russian girl named Olga Stempeliu.
4. At the age of 10, that is, in 1894 Rorschach was very much fascinated with Alfred Binet's idea of using inkblots for assessing the functioning of personality.
5. Rorschach started working on inkblots in 1911.
6. In 1912 Rorschach received an MD psychiatry under the supervision of Eugen Bleuler who was the teacher of CG Jung and had coined the term schizophrenia.
7. The title of PhD thesis of Rorschach was 'on Reflex Hallucinations and Kindred Manifestations'.
8. Based on his doctoral research, Rorschach published a research paper entitled 'Reflex Hallucination and Symbolism with Psychoanalytic Framework' in 1912.
9. In between 1912–14, Rorschach published many research papers on various clinical issues.
10. Rorschach also encouraged Pet therapy where he used to bring his pet monkey to the mental hospital for studying his patient's reactions while they were being entertained by the monkey.
11. Rorschach's most important work on inkblot test *Psychodiagnostik* was published in 1921.
12. On 2 April 1922, Rorschach died of appendicitis at the age of 37.

The entire procedure of the Rorschach test may be presented under the following three headings: *administration, scoring* and *interpretation*.

Administration

The administration of the Rorschach test is conveniently divided into three stages—first stage which is known as the *performance proper*, the second stage which is known as the *inquiry* and the third stage known as the *testing-of-the-limits*. In the first stage, the examinee is asked to take his seat and then the examiner talks with him for a few minutes to establish rapport. Handing over the first card to the examinee with the top side up, the Rorschach's basic instruction "What might this be?" is given. Klopfer & Kelley (1942, 32) modified and extended the basic instruction of Rorschach as under: "People see all sorts of things in these inkblots; now tell me what you see, what it might be for you, what it makes you think of." The examiner carefully notes down the following events:

1. He notes down the reaction time, that is, the time which elapses between the presentation of the card and the examinee's first scorable score. The reaction time is symbolized as *t* (Semeonoff 1976). Exclamations and comments are not counted as scores. However, they are noted down by the examiner.

2. He systematically notes down the position of the card when the response is being given. When the card is kept upright, the symbol used for this is Λ; when the top is turned downward, the symbol becomes V; when the top is kept at the left side, the symbol used is <; when the top is kept at the right side, the symbol becomes >; and when the examinee rotates the card without stopping, the symbol used is (0). Some experts use no symbol at all when the top is kept upright.

3. The responses are recorded verbatim. There are two basic purposes for recording the responses in this way. First, it enables the examiner to read it clearly during the time of scoring and thus, it increases the precision of scoring. Second, it facilitates others to read the record so that they may know exactly what the examinee had said.

4. The examiner records the total time for which the subject keeps each card. This is symbolized by T (Semeonoff 1976). Since the examinee is free to see as many things as he can, this period is known as the *free association period* and the total time elapsed during the free association period of each card is noted.

When the examinee returns the first card, he is given the next card and so on for all the 10 cards. Thus, all the 10 cards are presented in consecutive order. The examiner should keep all the cards out of the reach of the subject and he must hand over each card to the examinee rather than place it before him. It is common during testing for the examinee to raise some questions. In administering this test, silence by the examiner is considered as very important by most Rorschachers. However, this can be interrupted when the examinee raises the questions. In answering these questions, the examiner must adopt a nondirect approach. A few examples illustrate the point:

Examinee: "How long should I keep the card?"
Examiner: "As long as you like".
Examinee: "Should I turn the card?"
Examiner: "It depends upon you." or "As you like."
Examinee: "Should I report more than one thing?"
Examiner: "Most people usually report more than one thing."

Thus the questions raised by the examinee are answered in a nondirect way which, in turn, provides encouragement to him. Sometimes, it is found that the examinee rejects the cards altogether without giving any response. If such rejection occurs on the first or the second card, the examiner should rethink his procedure or his choice of the Rorschach test as the appropriate test. However, if such rejection occurs on the eighth or the ninth card, the examiner should encourage the examinee by saying, "There is no hurry; try and see something in it." The examiner should see that sufficient time intervenes before the examinee rejects the card. Beck (1944) and Hertz (1969) suggest a minimum time of 2 minutes only after which rejection can be justified. Rorschachers differ with regard to the maximum time a card should be kept by the examinee. Beck (1945) suggests that the examinee can keep a card as long as 10 minutes whereas Piotrowski (1969) suggests five minutes as a reasonable maximum time for keeping a card. Sometimes, examinees (in case of extremely compulsive examinees) have been found to go on responding *endlessly* to a card. In such a situation, the examiner should interrupt. The total number of responses or *R* on a Rorschach card for most adults varies from 15 to 30. A deviation from this range may not necessarily indicate psychopathological traits but it should be considered very carefully. When the examinee gives a vague and uncertain response, the examiner should note it and it should be clarified in the second stage of administration, which immediately follows the first stage, that is, the 'performance proper'.

Inquiry is the second stage of administration of the Rorschach test. It usually follows when responses to all the ten cards have been obtained during the stage of the 'performance proper'. Rapaport, Gill & Schafer (1946) are, however, of the opinion that inquiry should be conducted immediately after responses to a card have been obtained and with the card placed out of sight from the examinees. Rapaport takes the plea that the examinee is apt to forget many associations or his thinking may be interrupted by presentation of subsequent cards. It is, therefore, essential that inquiry should be conducted immediately after the presentation of a card. Most Rorschachers, however, do not approve of Rapaport's viewpoint and continue to conduct an inquiry after the responses to all the ten cards have been obtained. The basic purpose of the inquiry is to provide all the information necessary for scoring the responses accurately. Therefore, the inquiry consists of asking the examinee questions about each of his responses in the presence of the location sheet so that it may be known what part of the blot is the basis of his response and why the particular part gives rise to the response. The purpose of keeping the location chart before the examinee is to make a permanent record of the area of the blot used by him in responding. Thus, the purpose of the inquiry is twofold. *First,* it allows the examiner to obtain further information regarding the examinee's responses and *second,* it also helps the examinee to clarify his responses by adding and expressing more about what had already been said. Ordinarily, the extent to which questions during the stage of inquiry are formulated are left entirely to the skill of the examiner. However, for success of the inquiry, the examiner should keep some basic principles in view. *First,* the examiner must frame appropriate questions when needed. Some responses are very clear-cut and need no questions by the examiner. Questions put unnecessarily during the inquiry would add nothing new. Therefore, questions should be asked only when the examiner is in doubt regarding its location or meaning. *Second,* the inquiry questions should be framed in a nondirect way. (Examples of nondirect questions have been given earlier.) *Third,* the inquiry questions should not be long. They should be precise and brief and must be formulated to evoke answers which are helpful in scoring.

Testing-of-the-limit is the third and final stage in the Rorschach administration. However, this stage is not needed for all examinees. In some instances, it is seen that examinees give responses which are uncommon and not ordinarily found in most of the Rorschach protocols. For example, the examinee might somehow be convinced that whole responses are better than part responses and therefore give only whole responses to all the ten cards. In such a situation, testing-of-the-limits is done through adequate encouragement by the examiner in order to see whether the examinee is capable of changing the response in order to evoke the responses found in most of the protocols. The stage is so-called because the procedure intends to test the limits of giving ordered responses on the Rorschach test.

Scoring

In the Rorschach test, 'scoring' refers to the classification of responses into different categories so that the product may reveal the personality of the examinee as a whole. Regrettably, Rorschach died before he could have conceptualized his scoring system. Therefore, it was left to his followers to complete the systematization of Rorschach scoring. Five American psychologists, namely, S Beck, M Hertz, B Klopfer, Z Piotrowski and D Rapaport tried their best and produced overlapping but independent approach to the test. It was found that nuances of scoring vary from one scoring method to another. Therefore, Exner and his colleagues have developed the *Comprehensive Scoring System* (CSS) by synthesizing their earlier approaches (Exner 1993, Exner & Winer 1995). The CSS agrees that from the point of view of scoring, there are four main categories of classification of responses.

1. Location
2. Determinants
3. Content
4. Popular Responses and Original Responses

A detailed discussion of the four categories is presented below.

Location

Location is the first and easiest system in Rorschach interpretation. Location refers to the part of the blot which produces a particular response. Whether the whole blot is producing a response or only a part of the whole blot is producing the response is indicated by this category of scoring.

In the former case, it is a whole response and the scoring is direct and simple but in the latter case, the scoring is further dependent upon the fact whether the used part is a commonly or rarely used one. A detailed discussion of the scoring based upon location of the responses along with its symbols is presented below.

The W response (whole responses): When the whole blot instigates the response, it is scored as *W*. Thus the criterion for the scoring of *W* is a sort of either–or phenomenon in which the examinee either uses the whole blot or does not. In general, a *W* response is considered a good response because it indicates an overall view of the entire situation. There are two subtypes of *W* responses: the cut-off whole symbolized as *W* and the confabulatory whole symbolized as *DW* or *Ddw*. The cut-off whole response, which was suggested by Klopfer & Kelley (1942), is one in which the examinee gives the response on the basis of the whole blot excluding minor details. For example, Card I may be perceived as 'Bat' excluding its lower projection or side projection. Cards I, II and III evoke frequently the cut-off whole response. Card II may be perceived as 'Buffalo' excluding the red colour whereas Card III may be perceived as 'Two women' excluding the red colour. Beck (1944) has, however, objected to the cut-off whole response and is of view that a response is either *W* or it is not. He argued that since there are no evidences of scorer reliability of the cut-off whole response and also because the literature does not make mention of any such response, it is difficult to accept it. The confabulatory whole or *DW* or *Ddw* response is one where "the examinee attends only to a detailed area of the blot, but then generalizes from that detail to the entire blot" (Exner 1974, 57). When the examinee attends to the usual detail, *Dw* is used and when the examinee attends to the unusual detail, *Ddw* is used. Such responses obviously indicate that the examinee has not distinguished the various areas of the blot to have a clear perception but instead, he has generalized from a single minor detail to the whole blot. The *DW* or *Ddw* response, therefore, indicates perceptual cognitive impairment. However, such responses on the Rorschach test are very unusual.

The D response (common details): The *D* response stands for a response based upon the 'usual' or 'normal' (sometimes called 'large usual') detail or part of the blot. The fundamental principle for a particular response to be rated as *D* is that a detail should be easily and frequently seen by most examinees as separate from the remainder of the blot. It should not be confused with a 'popular response' of the content category (to be discussed later on) because the latter is different from the former. Rapaport, Gill & Schafer (1946) have defined the *D* response as belonging to the area "which is conspicuous by its size, its location and the frequency of the response it draws." Thus the 'normal' or 'usual' details of the blot are the areas which clearly stand out by virtue of their position and contour. A low percentage of *D* response is usually indicative of maladjustment. Klopfer et al. (1954) have divided the usual or common details into two parts depending upon the size—capital *D* for large usual area and small *d* for small usual area.

The Dd response (unusual details): The unusual detail responses (also called small detail responses by Rorschach) are based upon "usually the smallest details of the picture almost always overlooked by normal subjects" (Rorschach 1942). In the words of Exner (1974, 56), a *Dd* response is "one given to a blot area which is not used frequently." Thus when that part of the blot which is rarely used by most examinees instigates a response, it is scored as *Dd*. Logically, a response can be scored as *Dd* when it is neither *W* nor *D*. Klopfer and Kelley (1942) have divided the *Dd* response into four subcategories and the symbols for each are different. When tiny areas evoke a response, it is scored as *dd*; when an edge or contour instigates a response, it is scored as *de*; when inside areas evoke a response, it is scored as *di*; and when an unusual detail response does not fall into any of these categories, it is scored for rare detail as *dr*. A *dr* response is highly idiosyncratic and rare, and sometimes the examiner fails to see in the way the examinee perceives. Exner (1974) has, however, objected to Klopfer's division of the unusual details on three grounds. *First*, the examiners tend to disagree with specific types of *Dd*. For example, a response can both be tiny and rare and therefore, can be scored as *dd* as well as *dr*. Likewise, some inside detail responses are not uncommon whereas others are extremely rare and uncommon. The former may be scored as *di* and the latter may be scored as *dr* or both types of response may simply be scored as *di*. Thus, there is room for the scorer to vary and this unnecessarily lowers the reliability of the four categories of Klopfer's division. *Second*, there are no other evidences for such a fine division of the *Dd* response. The limited research base does not make such a differentiation regarding the type of *Dd*. *Third*, the frequency of any of these categories of *Dd* response is very low and contributes very little to the scoring summary. Keeping in view what has been said above, Exner (1974) recommends that all types of *Dd* responses should simply be scored as *Dd*. This would decrease the confusion regarding the differentiations among scorers and therefore, would tend to increase the scorer reliability.

The S response (white space details): When the white space area is used as the basis of response, it is scored as *S*. The *S* response is never scored alone. According to Exner (1974, 57), it is "always used in conjunction with one of the three primary location scores such as W_S, D_S or Dd_S." Rorschach himself regarded *S* as one form of unusual detail, that is, *Dd*. Beck has also expressed a similar view and said that *S* responses are a subclass of *D* or *Dd* so that his symbols become *DS* or *Dds*. The chief rationale for using *S* in conjunction with other location scores is to maintain consistency in the evaluation of the three main location scores—*W, D* and *Dd*. Some Rorschachers have, however, attempted to modify Rorschach's original idea. They have suggested that the space response may be scored separately as *S* if the basis of response is only the white space area. But when space area is used in combination with other parts of the blot, it must be scored as *WS, DS* or *Dds* depending upon the nature of the response. When the white space area is scored separately, it is included in *W*, *D* or *Dd*. Hertz (1970) is one such Rorschacher who recommends separate scoring for white responses. Accordingly, she has distinguished between two types of white space responses—the common and frequently used white space detail scored as capital *S* and the rare and infrequently used white space detail scored as small s.

The DdD response (confabulated details): The confabulated detail response symbolized as *DdD* is scored in much the same way as *DW* and *Ddw*. According to Exner (1974, 60), a *DdD* response is "one in which the subject interprets an unusual detail of the blot and then generalizes from that interpretation to a common detail area." The basic principle for a response to be scored as *DdD* is that a common or usual detail area is interpreted secondarily and the primary emphasis is based upon an unusual detail area. As compared to *DW* and *Ddw*, the *DdD* response occurs much less frequently in a normal Rorschach protocol.

These are the basic location scores on the Rorschach test. Out of these, the first three location scores, namely, *W, D* and *Dd* are the primary ones: According to Nunnally (1970, 30), an average adult gives *6W, 20D* and *4Dd* responses.

A summary showing the symbols, their descriptions and the criteria of the Rorschach location scores is given in Table 11.1.

Table 11.1 Symbols, descriptions and criteria of Rorschach location scores

Symbol	Description	Criterion
W	Whole response	When all portions of the blot are used in giving a response
DW/Ddw	Confabulated whole response	When secondary emphasis is given to the interpretation of the blot as a whole and the primary emphasis is on a detail portion, either usual or unusual of the blot
D	Usual or common detail response	When response is given on the basis of frequently or commonly identified area of the blot
Dd	Unusual or uncommon detail response	When response is given on the basis of infrequently or uncommonly identified area of the blot
S	White space response	When response is given on the basis of white space area of the blot (usually scored in combination with *W*, *D* and *Dd*)
DdD	Confabulated detail response	When secondary emphasis is given upon a usual detail area and primary emphasis is given upon an uncommon or unusual detail area

Determinants

Determinants are the second most important phase of scoring in the Rorschach test. This scoring system is the most complex of all phases. Determinants refer to the features of the blot which have produced the particular response. Originally, Rorschach (1921) suggested five symbols for the scoring of determinants: *F* for form, *M* for human development, and three for colour responses—*FC* (for form-colour), *CF* (for colour-form), and *C* (for pure colour response). It is obvious that Rorschach's original system for scoring the determinants was very simple. His original system did not contain any symbol for scoring shading or *chiaroscuro* responses because Rorschach's original cards upon which his basic researches were based contained no variations in hue or colour. However, the shading features among the cards were created by a printing error and immediately after seeing the printed cards, Rorschach thought of the possibility of this new dimension (Ellenberger 1974). Accordingly, he added a sixth symbol for a shading 'chiaroscuro' response, which was posthumously published in his last research paper in 1923. The symbol was parenthesized *(C)*. Many Rorschach systematizers who worked after the death of Rorschach have conducted a series of comprehensive researches and formulated their own scoring symbols for response determinants depending upon the several characteristics of blots. All these characteristics descriptively fall into three categories: *(a)* features relating to form; *(b)* features relating to colour, and *(c)* features relating to shading. Despite this similarity among the different Rorschach systematizers like Mons, Beck, Rapaport, Klopfer, Exner, Hertz, etc., confusion persists probably because the symbols used show little agreement among them. Not only this, where the same symbols have been used by two or more Rorschachers, the criteria for use of the symbols differ to a great extent. The net result is confusion among the scorers. Exner (1974, 70) has divided the determinant categories into nine separate parts, for which altogether 24 symbols have been introduced. It should be borne in mind that the success of the determinants category depends upon proper inquiry. However, spontaneity should be the guideline for giving all determinants symbols. If there is reason to believe that the response has been given only after the question put by the examiner during the inquiry, it should not be scored as such. A detailed description of these nine categories of determinants is given below.

Form Determinant (F)

The form determinant symbolized by the capital letter F is meant for the response which has occurred purely because of the form or shape of the blot. Thus F is reserved for the response where nothing other than shape or form has contributed to the percept. Often, Rorschachers have found that the percentage of F in many protocols is high. F is relatively easy to score when the inquiry (done carefully) reveals that it is only the shape, which has evoked the response. For example, Card I may be wholly perceived as 'Bat' and on inquiry if it is revealed that this is because the two side projections appear as the two wings and the middle part looks like the centre of the body, the scorer, without fail, should score it as F. Some Rorschach systematizers have divided F into two categories—$F+$ and $F-$. When the form is clear and good, it is scored as $F+$ and when the form is vague, it is scored as $F-$.

Movement Determinants (M, FM and m)

Three types of movement responses are usually reported in the Rorschach test—human movement, animal movement and inanimate movement. Originally, Rorschach had recommended scoring for only one type of movement, that is, for human movement, because in his opinion animal movement and inanimate movements are in no way different from human movement in the basic sense. He, therefore, provided symbol M for such movements and no other symbols were provided. But later, Klopfer et al. (1954), Hertz (1970) and Piotrowski (1957) disagreed with Rorschach and provided separate symbols for scoring of animal movement and inanimate movement. Rorschach characterized M responses as being marked by 'flexion' or 'extension' depending upon the use of the centre axis of the blot area in the response. Thus, a movement which indicates 'pulling inside' the centre axis is the 'flexor', and a movement which indicates 'pulling away' from the centre axis of the blot area is the 'extensor'. Beck et al. (1961) have, however, added a third type of movement response because in their opinion Rorschach's flexion–extension does not explain all movements. This type of movement, is the movement, which is 'static'.

The symbol M is, thus, reserved for human movement. Although Beck, Klopfer and Hertz accept this original derivation from Rorschach in toto, Rapaport, Gill & Schafer (1946) have accepted the same with some modifications. According to Rapaport and his colleagues, M should be restricted to only those responses which contain human figures—either complete or incomplete. Exner (1974, 74) has also accepted the symbol M for human movement but with some variations. On the lines of the flexion–extension division of Rorschach, he divides human movement into two parts—active and passive. Running, walking, jumping and crying are all examples of active human movement and sleeping, talking, smiling, and thinking are examples of passive human movement. For active movement, he recommends the addition of a superscript a for active movement and p for passive movement so that the complete symbol becomes M^a and M^p respectively. There are some *precautions* in scoring for M.

First, F should not be used when M is scored because M automatically assumes form or shape.

Second, when there is obvious reason to believe that a response of human movement is not spontaneous rather than provoked by the inquiry of the examiner, M should not be scored. (This is held to be true for all determinant symbols.) For example, when the examinee responds to Card VII with "It looks like two women", this would not be scored as M because the response appears to have been given only on the basis of form *(F)*. But during inquiry when the examiner asks "Are they doing something?" and if this is answered by the examinee as "Yes, the two women are talking" then such response should never be scored as M because the response is not spontaneous.

Third, sometimes *M* is scored also for the response where content is animal but the movement involved is found only among humans and never among animals. For example, a response like "Two butterflies talking with each other" to Card III and "Two pigs playing football" to Card VIII would be scored simply as *M*.

Animal movement is scored as *FM*, which was originally suggested by Klopfer in 1936 and frequently used by Piotrowski and Hertz. Rorschach himself did not attach much significance to such responses. No other Rorschachers except Piotrowski, Hertz and Klopfer have used the scoring for *FM*. All responses involving animal movement such as "Two bears going somewhere", "Two cats fighting", "The rat is jumping", "The buffalo is going" are simply scored as *FM*. Like human movement, animal movement may be active or passive. Accordingly, the superscript *a* for active movement and *p* for passive movement should be added to *FM*. Precaution should be taken in scoring *FM* for only those responses of animal movement which are spontaneous and not provoked by the examiner's inquiry or questioning.

Inanimate or inorganic or insensate objects producing a movement are scored as *m* originally suggested by Piotrowski in 1936 and incorporated by Klopfer and Hertz separately in this system. No other Rorschachers have used the scoring for *m*. For all types of inanimate movements, Piotrowski recommended the scoring of *m*. Klopfer and Hertz, however, disagreed with Piotrowski and suggested a combination of the symbol *F* with *m*. When form or shape dominates in the perception of an inanimate movement, it should be scored as *Fm* and when movement dominates over form, the symbol should be *mF* and where there is pure movement, it should be simply scored as *m*. But these three scoring categories—*Fm*, *mF* and *m*—did not prove to be popular because they overlapped with *FM* or *M* categories. Here Piotrowski's suggestion has been very popular, according to which all types of inanimate movements should simply be scored as *m*. Some inanimate movements may be active and some may be passive. Examples of active movement are "Aeroplane flying speedily", "Top spinning fast" and "Car going speedily" and for such movements, the superscript *a* should be added so that the complete symbol becomes m^a. Examples of passive movements are 'Ship moving slowly', 'Train crossing the bridge at a slow speed' and 'Moon crossing a cloud slowly' and for such movements the superscript *P* should be added to obtain the complete symbol, m^P.

Colour Determinants: Chromatic (C, CF, FC and Cn)

Originally, Rorschach divided the colour responses having chromatic features given to 5 cards (Cards II, III, VIII, IX and X) into three categories: (i) those responses which are wholly based upon colour and are scored as *C*; (ii) those responses where colour dominates but also the form or shape is involved and are scored as *CF*; and (iii) those responses where form dominates but colour is also a determinant. Such responses are scored as *FC*. Thus colour responses based upon chromatic features are *C*, *CF* and *FC* and probably these are the only scoring symbol in a Rorschach test where there exists the least disagreement, among the Rorschachers. Rorschacher also added a fourth symbol, *CC*, for colour naming, which was later modified by Piotrowski as *Cn*. A detailed description of these colour scoring determinants is as follows:

C is scored for pure colour response, that is, for responses where other factors except colour are irrelevant. A few examples of pure colour response are blood, ice cream, paint, etc. Pure colour responses are, however, very rare in the Rorschach protocols. *C* or colour response is interpreted to indicate *lack of control over emotional responsiveness*.

CF is scored for the response where colour features of the blot are primary and form features are secondary. Thus the responses where colour is dominant and form or shape is vague and undifferentiated are usually scored as *CF*. A few examples of *CF* responses would be fire, lake meat, water, scenes, sky, etc. Scoring of *CF* is often a difficult task for the scorer because some of these responses may be scored as *FC*, too. For example, when the examinee emphasizes that he perceives "meat" because it looks like a piece of meat, then obviously it would be scored as *FC* rather than *CF* because here form is primary and colour is secondary. Thus, through a nondirect inquiry the examiner must get himself fully satisfied that the percept has primary reference to colour and secondary reference to form in order to score it as *CF.*

FC is scored for a response where form dominates and the examinee also makes a secondary reference to the colour for the purpose of elaboration or classification. A few examples of an *FC* response would be lungs, heart, a red dog, a red shirt, a brinjal, etc. In all these examples, the form of the percept dominates with a secondary emphasis on the colour. A part of the blot may look like a heart because it is first similar in shape and then, it has red colour. However, a skilled examiner should get this cleared during the inquiry. Even after that if the examiner is in doubt regarding whether or not form dominates over colour, testing-of-the-limit may be carried out. As compared to *CF, FC* represents a very controlled use of colour. *FC* indicates *control over emotional impulses* whereas *CF* indicates *comparatively little control over it.*

Sometimes it is found that the examinee simply names a particular colour and declines to comment further on it. Such responses in which only the colour had been named were recognized as colour-naming responses and were scored by Rorschach as *CC*. However, he did not attach much significance to a *CC* response. Piotrowski, for the first time in 1936, introduced the symbol *Cn* in place of *CC* and this symbol has been frequently used by many Rorschachers. A few examples of colour-naming responses are "Here it is pink", "This is a red colour", "Greenish", etc. Exclamations like "Look at the colours!" or "What a nice variety of colours!" are not scored as *Cn*.

Colour Determinants: Achromatic (C', C' F and FC')

Black, white and grey are regarded as achromatic colours. Although these colours are not regarded as colours by a physicist or even by a psychologist, they are important because they are frequently used in day-to-day life. Rorschach himself never suggested a separate scoring for these achromatic colours. However, this was introduced for the first time by Klopfer in 1938 who based the decision to score achromatic colours separately on researches done subsequent to Rorschach as well as upon the viewpoints expressed by Binder in 1932. Rapaport included the same scoring symbols and criteria in his system whereas Hertz and Piotrowski modified the scoring symbols but the purpose and the criterion were those of Klopfer.

Achromatic responses, on the pattern of chromatic responses, fall into three categories and are symbolized by three different symbols: (i) the pure achromatic colour response scored as *C'*, (ii) the achromatic colour-form response scored as *C'F*, and (iii) the form-achromatic colour response scored as *FC'*.

The *C'* response is a pure achromatic colour response and has no reference of form or shape at all. Such a response is very rare and uncommon. A few examples of pure achromatic colour response are 'coal', 'mud', 'soil', 'snow', etc., and these are scored as *C'*.

The achromatic colour-form response is scored as *C'F* and refers to a response in which achromatic colours are the primary determinant and the form has a vague and undifferentiated reference. A few examples of *C'F* response are 'Black cloud', 'White water', 'Gray sky', etc. A part of the blot may be perceived as 'Black cloud' most probably because the cloud is blackish first, and then because its shape resembles a cloud. A skilled examiner, however, must confirm this from the examinee through inquiry before he scores it as *C'F*.

F'C represents the form-achromatic colour response in which the primary emphasis is based upon the form and the achromatic colour has only a secondary reference for the purpose of elaboration or classification. The decision to score a response as *F'C* is relatively easier than *C'F* because of the primary emphasis upon the form features. A few examples of the *F'C* response are 'Black dog', 'White horse', 'Gray cow', etc. Obviously, a part of the blot may be perceived as black dog because the blot looks like a dog first (emphasis upon form) and then because it is black (a secondary reference to colour). However, the examiner should confirm this from the inquiry.

Shading Determinants

Perhaps the least researched and the most controversial determinant category in the Rorschach protocol is that of shading or *chiaroscuro* responses. A shading response is one is which light-dark features of the blot are used. As earlier mentioned, Rorschach originally recommended only five symbols for determinant categories in which no mention of shading responses was made. But later, he also included a parenthesized (*C*) as a sixth symbol for shading responses. Since then, some of the Rorschachers have worked over the determinant category of shading responses and the net product has been the emergence of a different set of symbols with different criteria.

Beck has divided shading determinants into three categories. The *first* category includes those determinants in which shading produces the impression of texture and depending upon the dominance of form, this category is scored as *T, TF* or *FT* for pure texture or reflections, texture-form and form-texture responses respectively. The *second* category includes those determinants in which shading creates the impression of depth or distance, and again depending upon the dominance of form, they are scored as *V, VF,* or *FV* for pure vista response, vista-form response and form-vista response respectively. The *third* category includes determinants in which responses are based upon shading features (light–dark features) in which achromatic colours are also involved. Depending upon form involvement, such determinants are scored *Y, YF* or *FY* for pure shading response, shading-form response and form-shading response respectively.

Likewise, in 1936, Klopfer divided the shading determinants into *four* categories. The *first* category, which includes shading responses that tend to produce the impression of texture, are scored as *C, cF* or *Fc* for pure texture response, texture-form response and form-texture response respectively. The *second* category includes those determinants in which shading is perceived as general-diffuse. The symbols used are *K* for shading responses perceived purely as diffuse and *KF* for the diffuse form shading response. The *third* category involves those responses in which shading is used for vista (depth or distance), landscapes, reflections, etc. The determinant symbol for such responses is *FK*. The general demerit of the Klopfer system is that it fails to provide a clear distinction between *FK* responses and those involving a more diffused use of shading (scored as *KF* and *K*). The *fourth* category which was added by Klopfer in 1937, includes X-ray or topographical map responses and depending upon the dominance of form they are scored as *K*, for pure X-ray or topographical map responses, *KF* for X-ray form response, *FK* for form X-ray responses. Likewise, Hertz also used three categories of shading determinants in addition to separate scoring for the achromatic colours. Rapaport, however, used only two categories of scoring determinants in addition to separate scoring for the achromatic colours. Thus, shading determinants represent the widest differences among the Rorschach authorities. Despite these differences, Exner (1974) has been able to prepare a very comprehensive and widely-accepted scoring for the shading determinants, which utilizes the three basic symbols of Beck—*T, V* and *Y* for the three categories of shading responses. Exner's comprehensive system has, however, modified the criteria for two of the three categories of shading responses (that is, for *V* and *Y*) from those used in the Beck system. A detailed discussion of *T, V* and *Y* categories is presented below.

Texture Determinants (T, TF and FT): Klopfer was the first person to include a separate symbol for texture responses. His symbol was (*C*). Hertz later included Klopfer's symbol in her system. Piotrowski (1957) does not provide a separate scoring for texture responses. Rapaport et al. (1946) scored texture responses only when they were given in combination with achromatic colour responses. It was Beck (1944) who modified the symbol of Klopfer as *T* to represent texture responses and since then, texture responses are usually reported in the form of *T* by most Rorschachers.

Texture is scored when shading (light-dark) features of the blot are used to represent tactual stimuli. The common features of tactual stimuli are cold, hot, rough, hard, soft, smooth, sticky, greasy, furry, silky, etc. When these words are used by the examinee, it is highly probable that shading is involved and texture should be scored. Texture is scored in one of the three ways depending upon the extent of form involvement. *T* is scored for a response in which shading features are represented as tactual or textual without the involvement of form or shape. In other words, *T* is scored for pure texture response. Responses like 'Silk', 'Ice', 'Hair', 'Flesh' and 'Wood' may be scored as *T* if shading (light-dark) features of the blot are involved and perceived as texture with no form involvement. A pure texture response is, however, very uncommon among the three texture responses.

TF is the second category of texture scoring. A response in which shading is perceived as texture and form is involved only secondarily, is scored as *TF*. In other words, *TF* is scored for a response in which interpretation of shading as texture is primary and form features are used secondarily for the express purpose of elaboration or classification of the percept. Responses like 'A soft piece of ice', 'A hard metal', 'A rough skin',. 'A rough piece of sandpaper' are likely to be scored as *TF* provided shading features of the blot are involved.

FT is scored for the form-texture response, that is, for a response in which form is primary and the shading features of the blot perceived as texture are secondarily involved. Thus, in *FT* response the primary determinant is form (that is, form is distinct and clear) and the secondary determinant for the purpose of elaboration or classification is the shading features perceived as texture. For example, responses like 'Fur coat', 'A glass made of hard metal', and 'Smooth chair' are likely to be scored as *FT* provided shading (light-dark) feature of the blot are involved. In all these examples the form is clear and distinct and hence, it is the primary determinant.

Researches have revealed that of the three texture responses, three fourth responses are FT and occur with great frequency in all Rorschach's Cards except IV and VI (Exner 1974).

Shading-dimensionality or Vista Determinants (V, VF and FV): When shading or the light-dark features of the blot are perceived as representing depth and/or dimensionality, they refer to the shading-dimensionality or vista determinants. Rorschach made only a passing reference to such shading responses, which involved dimensionality. However, it was Klopfer who for the first time provided a separate scoring symbol (*FK*) for these responses in 1942. But unfortunately this symbol could not be popular as it was not clearly differentiated from *KF* and *K* responses, which were also introduced by Klopfer for responses involving the more diffuse use of shading. Following Klopfer's lead, Beck for the first time used for these vista responses the more distinguishing symbols—*V*, *VF* and *FV*. Like texture response, vista response is also scored in one of three ways depending upon the extent of form involvement. Beck included reflections or the responses based upon the reflections under vista responses and scored them as *FV*. Klopfer has also included such responses, for which the symbol *FK* has been provided in his system. Exner (1974) has not included the reflections under the vista response and has provided a separate scoring category for such responses by providing a very distinguishing status to them. The author, therefore, accepts Exner's view to be a more appropriate one.

V is scored for pure vista response. A pure vista response is one in which shading features are perceived as representing depth or dimensionality without any form movement. A few examples of the pure vista response are 'Deepness', and 'Height', provided the shading features are involved. Pure vista responses are usually very rare.

VF is the second scoring determinant for shading responses having depth or dimensionality. A response is scored as *VF* if the primary determinant is the shading feature perceived as depth and the secondary determinant is form or shape. In other words when a response utilizes shading features of the blot to represent depth and also includes a secondary emphasis on form for the purpose of elaboration or clarification, it is scored as *VF*. Responses like 'Mountain peak', 'Depth of a lake', and 'An aerial view of the city' may be scored as *VF* provided shading features have been used.

FV is the third scoring determinant for the vista response. A response in which the primary determinant is form and the examinee provides shading features of the blot as depth only as a secondary interpretation, is scored as *FV*. Thus in *FV* responses, form is dominant and primary and the shading features perceived as depth or dimensionality are vague and secondary. Responses like 'A deep well', 'A tall tree' and 'A woman behind a curtain' may be scored as *FV* provided shading features of the blot are involved. In all these examples, form is distinct and clear and is the base of the response. If the inquiry fails to satisfy the examiner regarding the dominance of 'form' or depth and/or dimensionality, testing-of-the-limits is recommended. Of the three vista responses, *FV* occurs in greater frequency. Following Beck, reflections or responses based upon reflections such as 'Two women in a mirror', 'A human face in the water of the well' would be scored as *FV*.

General-Diffuse Shading Determinants (Y, FY, and YF) The general-diffuse shading response includes within itself all those shading responses which are neither texture nor vista. As a matter of fact, it is this shading response to which Rorschach used the symbol parenthesized (*C*). Such shading responses are used in a more general and nonspecific sense than those of the texture or vista responses. Like the texture or vista response, the general-diffuse shading response can also be scored in any of three ways depending upon the degree of the form involvement.

A pure shading (general-diffuse) response is scored as *Y* and is defined as a response, which is exclusively based upon shading features whose content has no form. These shading features, however, must not be perceived either as texture or as vista. Responses like 'Darkness', 'Ink', 'Smoke', 'Fog' are scored as *Y* provided shading features of the blot have been used.

The shading-form response is scored as *YF*. A *YF* response is a response in which shading features are the primary determinant and the form is a secondary determinant, that is, form is used only secondarily for the purpose of elaboration or classification. Usually, the content of a *YF* response has vague and nonspecific form but the light-dark features are important for the formation of the content. Responses like 'Some sort of X-ray' and 'Dark clouds', can be scored as *YF* if shading features have been used.

The form-shading response is one in which form is primary to the formation of the response and light-dark features are used secondarily for the purpose of elaboration or classification. Such responses are scored as *FY*. Responses like 'Dark pen', 'Dirty shirt', 'Dark ship' may be scored as *FY* if the examinee gives emphasis upon form.

Form-dimensional Response (FD)

The form-dimensional response scored as *FD* is different from the form-vista response scored as *FV*, though form dimensionality is involved in both the responses. In an *FV* response the basis of dimensionality is the light-dark (or shading) feature of the blot whereas in *FD* response the basis of dimensionality is the form interpreted by the size of the blot areas. *FD* is a new scoring category introduced by Exner (1974). According to him, *FD* is scored for "responses which include perspective or dimensionality based exclusively on form, interpreted by size or in relation to other blot areas" (Exner 1974, 99). Responses like 'A bird on the top of a tree', 'A woman lying down', 'A statue on the top of a hill', 'A cross on the top of a church', may be scored as *FD*.

Reflection Responses (rF and Fr) and Pair Response (2)

Exner (1974) provided a separate category for the reflection responses based upon the symmetry of the blot. Beck and Klopfer have both suggested the inclusion of reflection responses as one form of vista response. Hertz has scored reflections in two ways—reflections based upon shading as shading responses and reflections based on form as form responses. Exner's scoring for reflection responses is the most compressible as well as easily adaptable one. Depending upon the form involvement, he classified reflection responses into two categories as mentioned below.

The reflection-form response is one category which is scored as *rF*. An *rF* response is one in which symmetry features of the blot are primary to the formation of the response and the form is used secondarily for the purpose of further clarification. Usually, in an *rF* response, the form is vague and nonspecific. Responses like "Sky reflected in a pond" and "A reflection of something in a lake" may be scored as *rF*. The frequency of *rF* responses is very rare.

The form-reflection response is the second category of the reflection response and is scored as *Fr*. The *Fr* response is one in which form features of the blot are primary to the formation of response and the content is perceived as reflected due to symmetry of the blot. Thus in an *Fr* response the form is distinct and clear, which distinguishes it from an *rF* response. Responses like 'A girl in a mirror' and 'Own face reflected in water' may be scored as *Fr* because of the dominance of the form, which is relatively clear and distinct. For Beck, all the above types of reflection responses would simply be scored as *FV*, which is a kind of vista response. Similarly, Klopfer would score them as *FK* which is also a kind of vista response.

Like reflection responses, pair responses are also based upon the symmetry of the blot and are scored as parenthesized (*2*). A pair response excludes the form specificity or form quality. Thus when the symmetry of the blot evokes the perception of twoness or pair, the symbol (*2*) is used by Exner (1974). Responses like 'A pair of flags', 'A pair of shoes', 'A pair of crabs'. 'Two bears', and 'Two little girls, one on each side' may be scored as (*2*). However, the examiner should carefully note that the symbol (*2*) should never be used when a pair or the two objects are perceived as *reflected*.

Thus, it is obvious that following Exner (1974) determinants have been divided into nine categories such as form, movement, colour (chromatic), colour (achromatic), texture (shading), dimensionality or vista (shading), general-diffuse (shading), dimensionality (form) and reflections and pairs. There are altogether 24 scoring symbols used for them. For convenience, a list of these symbols along with appropriate response names and categories is presented in Table 11.2. Readers should note it carefully because sometimes two or more than two determinants may be combined in one response. Such responses are known as blend responses which will be discussed in detail along with Beck's Organizational response, or *Z*, while discussing the interpretation of Rorschach response.

Table 11.2 Symbols, category and description of the Rorschach determinants

Symbols	Category	Description
F	Form	Form response
M	Movement	Human movement response
FM		Animal movement response
m		Inanimate or inorganic movement response
C	Colour (chromatic)	Pure colour response
CF		Colour-form response
FC		Form-colour response
Cn		Colour-naming response
C'	Colour (achromatic)	Pure achromatic colour response
C'F		Achromatic colour-form response
FC'		Form-achromatic colour response
T	Texture (shading)	Pure texture response
TF		Texture-form response
FT		Form-texture response
V	Dimensionality Depth or Vista (shading)	Pure vista response
VF		Vista-form response
FV		Form-vista response
Y	General-diffuse (shading)	Pure shading response
YF		Shading-form response
FY		Form-shading response
FD	Dimensionality (based on form)	Form-based dimensional response
rF	Reflection and pairs	Reflection-form response
Fr		Form-reflection response
(2)		Pair response

Content

A final task in scoring responses on the Rorschach test is to select the appropriate content. Usually, this final task is very important for interpretation of the responses and that is done by appropriately selecting the symbols to represent the content. The selection of appropriate symbols to represent the content is not a very difficult task.

Originally, Rorschach in 1921 used only six symbols for the scoring of content. They were *H* (human), *Hd* (human detail), *A* (animal), *Ad* (animal detail), *Ls* (landscape) and *Obj* (inanimate objects). After Rorschach, several scientists found that these six categories did not provide an adequate distinction among the content categories. As such, they developed their own symbols based upon Rorschach's original symbols to represent the content categories in detail. The longest list consisting of 35 content categories was provided by Beck (1944), and the shortest list consisting of 23 content categories was provided by Klopfer & Davidson (1962). Exner (1974), in his comprehensive system, has included only 26 content categories out of which 22 are the basic content categories and four are the parenthesized supplementary content categories. Various

Rorschach systematizers not only give different lengths to the list of content categories but also use different symbols for the same category. For example, Klopfer has used symbol *At* for the anatomy content whereas Beck has used the symbol *An* for the same category. Since the lists of content categories proposed by most Rorschachers vary widely, it is difficult to prepare a list acceptable to all. However, the most common categories of content, which are acceptable to most Rorschachers, are reproduced below.

Human Responses H, (H), Hd, (Hd)

H is used for whole or nearly whole human figures. Responses like 'Man' 'Woman', or 'Man having no hair on the head' would be scored as *H*. (*H*) is used for mythological (or fictional) or cartoonlike whole human figures. Thus responses like 'Giant', 'Fairy', 'Ghosts', 'Dwarf' and 'Devils', etc., would be scored as (*H*) rather than *H*. The symbol *Hd* is used for parts of the human body or the incomplete human form. Limbs, head, feet, hands, fingers, person without head, etc., would be scored as *Hd*. Parts of the human body considered from the mythological or fictional point of view are scored as (*Hd*). Thus 'Eyes of the devil', 'Finger of a ghost', 'Leg of a monster' would be scored (*Hd*) rather than *Hd*.

Animal Responses A, (A), Ad, (Ad)

The symbol *A* is used for whole or nearly whole animal figures. Responses like 'Buffalo', 'Cow', or 'Cow having no horn' would be scored as *A*. When the response of a whole animal figure is given from the mythological or fictional point of view, the symbol (*A*) is used. Responses like 'Magic horse' and 'Flying fish', are scored as (*A*). *Ad* is reserved for parts of the animal body and (*Ad*) is used for parts of the animal body considered from the mythological or fictional angle. Thus 'Head of the buffalo', 'Tail of a dog', 'Head of a frog', and 'Ear of an ass' would be scored as *Ad* and 'Wing of magic horse' and 'Face of a lion of Goddess Durga' would be scored as (*Ad*) because of the mythological viewpoints expressed therein.

Anatomy (An)

An is used for anatomy (internal organs) of humans or animals. Thus responses like 'Lungs', 'Heart', 'Skull', 'Kidney', and 'Intestine' (of either human being or animal) would be scored as *An*. Klopfer has used *At* in place of *An* for anatomy content.

Nature (Na)

Natural objects or things are scored as *Na*. A few examples of the natural objects are rainbow, waterfall, storm, night, thunderstorm, fog, mist, and so on. All these responses would be scored as *Na*. Clouds, are, however, excluded from this content category because of its importance and a separate symbol *Cl* is provided (Exner 1974).

Botany (Bt)

Objects representing plant life such as flower, fruit, tree, bush, and so on are scored as *Bt*.

Blood (Bl)

The symbol *Bl* is used for the blood of either human beings or animals.

Art

Paintings, family, crest, and the seal of authority are written as *Art* in content scoring.

Clothing (Cg)

Clothing materials associated with human beings such as trousers, hat, blouse, sari and hooks are scored as *Cg*. Clothing materials associated with mythological characters such as a ghost's trousers or a witch's hat are scored as parenthesized (*Cg*).

Food (Fd)

All responses representing food substance or edible objects are scored as *Fd*. Responses like apple, fried chicken, eggs, fried fish or meat, etc., are scored as *Fd*.

Fire (Fi)

Responses like burning, electric bulb, burning candle, actual fire, flame coming out from a stove or torch, etc., will be scored as *Fi*. In the determinant category, these responses are scored as *m* because they denote inanimate movement. Responses like atomic explosion or blast are, however, given a separate content symbol, that is, *Ex*, which denotes explosion. The determinant category will continue to be *m*.

Household (Hh)

Interior household items such as table, plate, chair, bedsheets, pillow, etc., are scored as *Hh*.

Landscape (Ls)

Precepts involving landscapes or seascapes are scored as *Ls*. Thus responses like beautiful garden, an aerial view of the city, underwater scenes are scored as *Ls*.

Sex (Sx)

Responses representing sex organs or sexual functions are scored as *Sx*. Responses like penis, vagina, sexual intercourse, breast, testicles, menstruation, and so on are scored as *Sx*.

X-ray (Xy)

Responses involving percepts of X-ray are scored as *Xy*. Thus X-ray of the bones, X-ray of the heart, X-ray of the intestines, etc., are scored as *Xy*.

Popular and Original Responses

The popular or *P* responses are those responses which occur frequently in the Rorschach protocols. Originally, Rorschach (1942) made no mention of the popular responses in his work. However, he recognized the importance of such responses in his posthumously published paper in 1923. He called these popular responses *vulgar responses*. According to him, a *P* response indicates the ability for *conventional perception*. Rorschach set up an objective criterion in defining the *P* response. According to him any response to be recognized as a *P* response, should occur at least once in every three protocols. This one-in-three criterion of Rorschach has been adapted by most Rorschachers. Klopfer & Davidson (1962) for example, have developed a list of popular responses on the basis of their clinical experiences using the one-in-three criterion. Rapaport et al. (1946) defined the *P* response as that which occurs at least once in every four or five protocols. Piotrowski (1957) has recommended the scoring of *P* as responses occurring at least once in every four protocols. Hertz (1970) defined the *P* response as a response that occurs at least once in every six protocols, thus using the broadest criterion.

Setting aside some minor variations, among Rorschach authorities like Hertz, Mons, Piotrowski, Rapaport, Klopfer, Exner and Beck, the popular responses evoked by the ten cards may be enlisted as shown below.

Card I:	Bat or butterfly (*W*), human figure (*D*)
Card II:	Human figures (*W*), butterfly (*D*), animal forms, usually the heads of dogs or bears (*W* or *D*)
Card III:	Two human figures or single human figures, and dolls (*W*), butterfly (*D*), fish (*D*)
Card IV:	Animal skin or a human figure covered in animal skin (*W*), shoe or boot (*D*)
Card V:	Moth, vulture, eagle, bat and butterfly (*W*), rabbit (*D*), *leg (d)*

Card VI: Animal skin (*W*)
Card VII: Head or face of a woman or a child (*W* or *D*)
Card VIII: Animal figures commonly perceived as wolf, fox, coyote, dog and bear (*D*), tree, bush (*D*)
Card IX: A head or face of a male (*D*), camel's head (*D*), tree (*D*)
Card X: Many-legged creatures such as crab, lobster, spider and rabbit's head (*D*), dogs (*D*)

Rorschach (1942) had also suggested scoring for the original response. He defined the original or *O* response as a response that occurs no more than once in one hundred protocols. Therefore, original responses are rare and creative responses. On the basis of the form quality, the *O* response is divided into two categories. An original response having a clear and distinct form is scored as O^+, whereas O^- is scored for a vague and indistinct original response. Almost all Rorschachers except Beck have included the scoring of the original response in their system but each of them has provided a word of caution. Piotrowski (1957) has said that the scoring of the original response is highly subjective and should be accomplished only by some trained Rorschachers. Likewise, Rapaport et al. (1946) have suggested that the decision to score the original score is a very difficult task and only the trained Rorschacher should attempt it. Likewise, Klopfer & Kelley (1942) consider the scoring of the original response as "a hopeless enterprise" because the number of such responses is unlimited.

INTERPRETATION OF THE RORSCHACH PROTOCOL

The interpretation of Rorschach data is a complex task and requires considerable training, skill and experience. Perhaps, this is the reason why there is no complete consensus regarding the interpretation of the data even among the prominent Rorschach authorities. However, in this book an attempt has been made to present the interpretative significance of Rorschach data in the simplest possible form, which beginners may find profitable. Throughout the interpretation of Rorschach responses, the method of content analysis (also known as document analysis) has been assumed to be the correct method of analysis.

According to Rorschach, the total number of responses or *R* varies between 15 to 30 for a normal examinee. Klopfer, however, estimated *R* to vary between 20 and 45 and Exner (1974) estimated *R* to vary between 17 and 27 under his comprehensive system. Thus the minimum total *R* or the maximum total *R* varies. A minor deviation from the above range does not necessarily indicate psychopathological traits. However, it should be explained with caution. A very low total *R* such as 10 or below indicates rejection by the examinee, that is, he has rejected one or more cards of the test and the skilled examiner should make a careful note of it because it may indicate defensiveness, intellectual limitations, depression and organicity. However, none of these traits can be wholly accepted or rejected solely on the basis of the total *R*. Other factors to appear later should also be taken into account. Likewise, a very high total *R* such as 90 or 100 indicates *schizophrenic tendencies* (Piotrowski 1957).

Almost every Rorschacher has emphasized the reaction time of the scorable response in each card as well as the total time elapsed during the free association. Piotrowski (1957), however, remarked that Rorschach never used a stopwatch and he never recorded the reaction time with any precision, though he noted down a longer pause. The general purpose of noting down reaction time was to make a comparison of the average reaction time to the chromatic colour cards (Cards II, III, VIII, IX, and X) with that of the achromatic colour cards (Cards I, IV, V, VI and VII). Such comparison may reveal the presence or absence of what is called the *colour-shock*, which is indicated when the average reaction time to chromatic cards is higher than the average reaction time to achromatic cards.

A detailed discussion of the interpretative significance of the different symbols used in the four categories of scoring, that is, location, determinants, content as well as popular and original responses is presented below.

Location Scores

All location scores, in general, indicate how the examinee approaches his environment, particularly the ambiguities or vagueness pertaining to it without answering why. In other words, the location scores indicate only the manner in which the examinee approaches the ambiguities of his world without answering why he prefers this manner or that manner. A content analysis of the various location scores is presented below.

Of the various location scores, the *W* score is probably the most frequently and widely investigated location score. *W* responses are given in greater frequency for Cards I, IV, V, and VI because these cards are more solid and represent the form of unbroken blots. Rorschach (1942) had revealed that *W* has a direct relationship with intellectual ability and capacity to organize things or objects in a meaningful and coherent whole. A review of the literature reveals that this revelation of Rorschach has aroused some controversy because some researchers have failed to show a consistent relationship between *W* responses and mental ability. For example, Amirtage et al. (1955) reported a very low correlation between *W* responses and IQs. Likewise, Lotstaff (1953) reported that *W* responses are related to verbal fluency only but not to the overall intelligence. Wittenborn (1950) found no relationship between *W* responses and measures of intelligence. Because *W* responses are related to intelligence only for some specific cards, there is low correlation (about .40) between the number of *W* responses and IQ scores (Kaplan & Saccuzzo 2001). The proportion of *W* is also an important clue. Exner (1974) reported that the ratio of *W* to *D* (*W*:*D* ratio) for a normal adult should be 1:2. If the *D* side of the ratio is elevated, it is ordinarily concluded that the examinee selects the easier perceptual-cognitive mode to act when faced with ambiguity. On the other hand, when the *W* side of the ratio is elevated, it indicates that the examinee reports to excess in a bid to organize an ambiguous situation. *W* is also interpreted with reference to the human movement determinant, that is, *M*. The average normal adult ratio of *W* to *M*, that is, *W*:*M* (both indicative of intellect) is 2:1 to 3:1. If *W* exceeds the 3:1 ratio, it indicates that the examinee has a high aspiration level, which normally exceeds his ability level. On the other hand, if *W* falls below 2:1 it indicates that his aspirations are below the level of his ability.

The *D* response, according to Rorschach, indicates the ability to perceive and react to clear and distinct characteristics of the world. Generally *D* responses are easier to give than *W* responses because they represent the easiest perceptual cognitive mode to act when faced with ambiguity. A review of the literature reveals that Card X produces the largest number of *D* responses because of its unbroken features. Usually, in a protocol the number of *D* responses is higher because *D* responses are easier to give than other location responses. Persons having stress and anxiety give only a few *D* responses. Even maladjusted persons yield a low proportion of *D* responses.

The *Dd* response is interpreted as a form of respite from the vagueness of the common or larger areas of the blot. For a normal adult, the proportion of *Dd* as compared to *W* and *D* responses is very low. It occurs at the rate of 5% in the protocol of a normal adult, that is, one response out of 20 is *Dd* for a normal adult. When the amount of mental disturbance is high as in the case of schizophrenics or obsessive compulsive neurosis, the percentage of *Dd* may exceed even 30%. Mere presence of the *Dd* response should not be interpreted as being an unhealthy sign because when it is present in appropriate ratio along with *W* and *D*, it indicates both initiative capacity and the capacity to withdraw. The research findings revealed that the *Dd* response was positively related with internal adjustment and was negatively related with external adjustment.

The *S* response, according to Rorschach (1942), Beck (1945) and Rapaport et al. (1946) indicates negativism or oppositional features. Klopfer et al. (1954) has, however, interpreted the *S* response to be the indicative of 'constructive self-assertiveness' provided such responses do not occur disproportionately in the protocol. A review of the literature supports the above conclusions with a few contradictions. Bandura (1954) reported that the *S* responses are positively related with 'oppositional' or negativistic tendencies. Murray (1974) has, however, found no such relationship between these two variables. Rapaport et al. (1946) reported that paranoid patients showed the highest frequency of *S* responses.

Lastly, the three confabulatory types of answers, that is, the *DW*, the *DdW* and the *DdD* responses, are also important interpretative elements. These are a highly unique form of response and are very rare. According to Rorschach, such confabulatory types of location responses indicate intellectual constriction and distorted perception. All Rorschachers except Beck have accepted this view of Rorschach and have considered the confabulatory responses to be uniformly pathological ones. Beck (1945) has shown that these confabulatory types of responses are given by a normal adult specially when his intellectual capacity is superior. Following Beck, such confabulatory responses indicate only a logical cognition (Exner 1974).

Determinant Scores

Determinant scores indicate psychological action, which forms the basis of a response. In other words, they indicate what psychological action underlies a particular response. A content analysis of the various determinant scores is presented below.

The *F* response (or the pure form response) is the most common determinant scoring in the Rorschach test. According to Rorschach (1942), a pure form response is related to the examinee's thinking or reasoning and indicates the attention–concentration features of human thinking. In other words, it indicates adequate affective delay or control necessary for the occurrence of the form response. Beck (1945) and Klopfer et al. (1954) have agreed with Rorschach but with an additional caution. They caution that the *locus* of the pure form response is indicative of constriction or defense and emotional conflict may be present there. But through deliberate thought operations, the conflict is somehow controlled. According to Rapaport et al. (1946), the pure form responses indicate a process of formal reasoning and the ability of the examinee to direct his attention and thinking to the elements of control and his ability to make discriminating judgement. Such examinees also show regard for the standards of the environment. Most of the Rorschachers have compared the pure form responses with the nonpure form responses for extracting adequate meaning from the pure form responses. One such comparison is done through calculating $F\% = \frac{\Sigma F \times 100}{R}$. Recently, Beck (1945) has argued for another method of comparing pure form responses with nonpure form responses. His method provides for the Lambda index. One advantage of the Lambda index over *F*%, according to Beck, is that it avoids the problems involved in interpreting the percentages (in *F*%), specially when *R* varies. The Lambda index is calculated by dividing the sum of pure form determinants by nonpure form determinants. Hence,

$$\text{Lambda index} = \frac{\Sigma F}{\text{Sum of non-}F\text{ determinants}}$$

The range for an average normal adult lies from 0.50 to 1. When the Lambda index exceeds 1, it indicates excessive affective constraint and when it falls below 0.50, it indicates affective instability. Several research findings have revealed, in general, that when the examinee is in a state of defensiveness, the number of pure form responses is increased, thus supporting Beck and Klopfer's position. A low frequency of pure form response occurs when the examinees are instructed to respond very quickly because this deprives them of the delay necessary for the formation of the pure form answer (Hafner 1958). It has also been reported that intoxication tends to increase the frequency of pure form responses but the quality of such responses is reduced. Likewise, it has also been demonstrated that psychopaths tend to give fewer pure form responses than alcoholics.

The pure form response has also been interpreted in terms of its quality, that is, in terms of good form (F^+) or poor form (F^-). According to Beck (1945), the F^+ response indicates that the examinee has respect for reality whereas the F^- response indicates that he has no respect for the reality of the environment. Some researchers have demonstrated a relationship between F^+ responses and the intellect. Klopfer & Kelley (1942) reported a significant positive correlation between a lower percentage of F^+ responses and mental retardation. In one of their early studies, Beck & Molish (1967) reported a high correlation between the lower percentage of F^+ responses and limited intellectual endowment. F^+ responses also indicate the ability to deal with stress effectively (Goldberger 1974). Some Rorschachers have argued for the calculation of an $F^+\%$ which is equal to $\frac{\Sigma F^+ \times 100}{(\Sigma F^+) + (\Sigma F^-)}$. For a normal adult (who is neither disturbed nor of low intelligence), $F^+\%$ should be above 80%. $F^+\%$ exceeding 80% is taken as indicative of the fact that the person has firm control over his intellect and behaviour.

Movement determinants (*M*, *FM* and *m*) are the second important interpretive elements in the Rorschach test. Out of the above three movement determinants, *M* is the most widely investigated one. According to Rorschach, *M* responses indicate the phenomena of 'internalization', that is, they indicate the ability of the examinees to handle the more deliberate and sophisticated inner experience (marked by organization and reasoning) in a way that they can be controlled emotionally. Thus *M* responses indicate ideational type of thinking, which grows with cognitive maturation. Klopfer et al. (1954) and Hertz (1974) have argued that *M* responses indicate a functional relationship between the inner world and the external world of reality. Beck (1945) suggests that *M* responses indicate awareness towards the external world and reflect some conflicts or emotions which do not get obvious expression in the world of reality. Thus the three prominent Rorschachers, namely, Klopfer, Beck and Hertz, have agreed with Rorschach's view that *M* responses indicate the phenomenon of internationalization because all of them regard *M* responses as bridging the inner world with the reality of the external world. Piotrowski (1960) and Rapaport et al. (1946), while agreeing with Rorschach, have maintained a slightly different position. Piotrowski argues that *M* responses indicate a sort of deeply embedded and not easily modifiable tendency to maintain a uniform attitude in dealing with others. Rapaport holds that *M* responses indicate a type of delay of response occurring due to some cognitive functions in a bid to give a more deliberate response. Thus, *M* responses are indicative of a resistance to a natural or spontaneous response, the result of which may be either internalization or externalization. Recently, some researches have been conducted in a bid to correlate *M* responses to different overt and covert behaviours. Dana (1968) conducted one study in which *M* responses were examined in relation to six important constructs—fantasy, time, estimation, intellect, creativeness, and interpersonal relations. His observation was that *M* responses may be related to any or all of the above six behaviours. Several studies have confirmed the positive correlation between intelligence and *M* responses (Abrams, 1955; R Sommer & D Sommer 1958). Cocking, Dana & Dana (1969) have reported that *M* responses are related with intellect, fantasies, and time estimation. *M* responses have also been interprete in terms of other responses on the test, especially the colour-responses, that is, in terms of th ratio of sum *M* to the weighted sum *C*. A detailed discussion of this ratio appears later in thi chapter in the context of *Erlebnistypus (EB)* or the 'experience balance'—one of the five variable introduced by Rorschach in the interpretation of protocol.

FM response (or animal movement response) is another important determinant much emphasized by Klopfer, Hertz and Piotrowski. According to them, *FM* responses indicate 'impulsiveness' of a person, that is, the trait of a person in which he is more guided by need for immediate gain or gratification than for long-term goals. The trait of impulsiveness is more common among children and aged persons than among adults. Accordingly, it has been reported by Klopfer & Davidson (1962) that the frequency of *FM* as compared to *M* is high for aged persons and Ames (1960) reported that in child protocols, the frequency of *FM* exceeds the frequency of *M*. Some independent studies have been conducted to examine the relationship between the *FM* responses and the correlates of impulsiveness such as aggressiveness, distractibility, irresponsibility, defensiveness, etc. In her study, Hann (1964) demonstrated that *FM* responses are related with the different measures of defensiveness such as substitution where overt expression of impulse is found, and with rationalization, regression and internalization where the impulse does not get an overt behavioural expression. Sommer & Sommer (1958) have found that *FM* responses are significantly correlated with aggressiveness and assaultiveness. Thompson (1938) demonstrated that irresponsibility, aggressiveness and distractibility as measured by MMPI highly correlated with the preponderance of *FM* responses. Berryman (1961) in this study of creative artists found that *FM* responses were positively related to their level of productivity.

The *m* responses (or the inanimate movement response) indicate impulses or thoughts which are 'threatening' to the stability of the personality and are usually beyond the control of persons (Klopfer et al. 1954). Hertz (1942) and Piotrowski (1957) interpreted the *m* responses in slightly different ways. While they did not consider such responses as threatening, they took them as representing impulses or thoughts which are not properly integrated into the cognitive functions. No other Rorschachers, except the above three, have included the scoring for *m* in their systems. The most common interpretation of *m* responses among the three Rorschachers is that such responses are associated with frustration occurring due to imperfect interpersonal behaviour. A few researches regarding *m* responses have revealed something more than what has been said above. Majumdar & Roy (1962) have reported that *m* responses occur in greater frequency among juvenile delinquents. Neel (1960) in her study has demonstrated that *m* responses increase with an increase in inhibitions, tension and conflicts, commonly created by the failure to integrate needs with behaviour. The active-passive dimension of the three movement scores, *M*, *FM* and *m* is also of some interpretive significance. Passive movements of the three types, in general, are related to depression and active movements are taken to be indicative of character disturbances.

The chromatic colour determinants (*C CF, FC* and *Cn*) are of next important interpretive significance. According to Rorschach (1942), these chromatic colour responses indicate affectivity or emotional excitability. *C, Cn* and *CF* responses in which colour-responses are dominant, indicate affectivity which has no capability for adaptiveness. *FC* responses, in which the form rather than the colour is dominant, indicate affectivity with the capability for adaptiveness. This is known as Rorschach's *colour-affect hypothesis*. As interpreted above, this hypothesis holds that when the proportion FC to $CF + C$ (that is, $FC : CF + C$) is high, the examinees possess the degree of control (capability for adaptiveness) during the affective or emotional state but when the reverse is true, it indicates that the examinees do not possess the degree of control during the affective or emotional state. In calculating the above ratio, the frequency of *Cn* determinants is included within the *C* determinant. There are two controversial issues relating to Rorschach's colour-affect hypothesis. The first issue relates to the concept of 'colour-shock'. This concept was introduced by Rorschach himself and was defined as the startle or surprise reaction to colour blots. The existence of colour-shock was recognized on the basis of some indices like longer reaction time to chromatic colour Cards (II, III VIII, IX and X), inability to respond to the colour cards, exclamation, etc. Colour-shock indicates anxiety neurosis and seriously impaired ability to respond, which is produced by the emotional disequilibrium due to affect produced by colour. The second issue of colour-affect hypothesis is related to what Klopfer & Kelley (1942) have called 8–9–10% and Beck et al. (1961) has called the 'affective ratio' (abbreviated as *Afr*). The affective ratio is equal to $\Sigma R(\text{VIII} + \text{IX} + \text{X}) : \Sigma R(\text{I to VII})$. Klopfer has expressed the same view in terms of percentage. The range of the affective ratio for a normal adult falls between 0.55 and 0.75. The affective ratio below 0.55 indicates withdrawal or passivity towards affective stimulation and the affective ratio above 0.75 indicates that the person has an uncontrolled tendency to be caught up by affective stimulation. In other words, this issue is related to the proportional number of responses to the last three cards—VII, IX, X, which are fully chromatically coloured. Both Klopfer and Beck have agreed that when the percentage or proportion of responses or *R* to the last three cards is high, the examinees are considered as responsive to affective stimulation. But when the converse is true, it indicates that the examinees are withdrawn from affective stimulation. This simple revelation has, however, not been supported by researchers like Meyer (1951), Perlanan (1951), Dubrovner, Von Lackum and Jost (1950). Some researchers have made correlational studies of chromatic colour-responses with interesting and important revelations. Gardner (1951) has obtained a significant positive correlation between chromatically coloured responses and impulsiveness. Kobler & Steil (1953) have demonstrated that chromatic colour-responses are not related to all kinds of affects. For example, in depression such responses tend to almost disappear.

The colour-naming response (or *Cn*) is very rare among chromatically coloured responses. Weigel (1941) has demonstrated that *Cn* indicates brain damage. Piotrowski has reported that *Cn* indicates some kind of organic involvement and thus has supported Weigl. It has also been found that *Cn* responses were given by persons with diencephelic lesions.

The achromatic colour responses (*C'*, *FC'*, and *C'F*), based upon the white-grey-black features of the blot, are of next interpretive significance. The *C'* response indicates the affective state in which the affect is apparently released but the person is hesitant to express it directly to the external world, that is, the affect is not exposed directly to the world (Klopfer & Kelley 1942). Since the affect is not exposed directly, its behavioural consequences are not so obvious as when the affect is not suppressed (usually reflected through chromatically coloured cards). The most common experiences associated with *C'* responses are tension, constraint and/or pain, all or any of which may lead to disequilibrium to cognitive stability. Rapaport et al. (1946) considered the *C'* responses to be indicative of conscious and deliberate control or defense against affective expression to the external world. Piotrowski (1957) has agreed with Rapaport and Klopfer in the interpretation of *C'* responses. However, he has added that such responses may also indicate euphoric character, especially when white or light gray features of the blot become the basis of such a response. Klopfer, Rapaport and Piotrowski also agree that when the basis of *C'* responses are black and grey features and black-grey-white features, they tend to indicate depression. When form dominates over *C'* responses (scored as *FC'*), it indicates that the impact of affect on cognitive stability is less intense and well-controlled, but when form plays a secondary role and *C'* dominates (scored as *C'F*), it indicates that the impact of affect on cognitive stability is disruptive and uncontrolled.

The shading responses are the next important determinants and hence of next important interpretive significance. As discussed earlier, shading responses are of three types—texture (*T*, *TF* and *FT*), vista (*V*, *VF*, *FV*) and general-diffuse (*Y*, *YF* and *FY*). Of these three, the texture determinants are the most popular ones. According to Hertz, texture responses indicate cautiousness in establishing affectively rewarding relationships with the external environment. Klopfer et al. (1954) have argued that the texture responses represent affective need. Beck (1945) suggests that texture responses indicate painful affective experience, which is usually related to infantile needs, particularly the erotic needs of the infantile period. When form dominates the texture responses (scored as *FT*), it indicates that the person is able to control his affective needs and can use them to his own advantage (Beck 1945). Klopfer et al. (1954) agree with Beck and have further suggested that *FT* reponses indicate "the awareness and differentiation of a persons's need for affection and dependency." Where texture dominates so that form or shape plays only a secondary role (scored as *FT*), it indicates that the painful affective experience is disruptive and interfering with proper interpersonal behaviour (Klopfer et al. 1954; Beck 1945).

The vista responses (symbolized as *V*, *VF* and *FV* by Beck) are the shading responses based upon dimensionality. Such responses are very rare and their absence from the protocol, in general, is regarded as a healthy sign (Exner 1974). Beck (1944) suggests that vista responses indicate a painful feeling tone in which depression of affect and inferiority feelings are also involved. Klopfer et al. (1954) interpret shading responses based upon dimensionality (or vista responses) as representing the persons's ability to deal with anxieties and conflicts objectively. There are a limited number of studies on behavioural correlates of vista responses and a few of them have revealed some important facts. Meltzer (1944) has demonstrated that stutterers give more vista responses than nonstutterers. Likewise, vista answers occur in greater frequency among alcoholics than among psychopaths (Exner 1974).

The diffuse-shading responses (scored as *Y*, *YF* and *FY* by Beck) are the third category of shading responses and are more common and frequent than the texture or vista responses in a protocol. The diffuse-shading responses, in general, indicate anxiety and withdrawal from the environment or passivity. When diffuse-shading responses are wholly related to anxiety, it is commonly referred to as shading-anxiety hypothesis, and when diffuse-shading responses are wholly related to the tendency for withdrawal, it is referred to as shading-passivity hypothesis. Klopfer et al. (1954) argue that such responses indicate a free-floating anxiety against which the individual remains unable to build any defences. He, thus, supports the shading-anxiety hypothesis. Rapaport et al. (1946) have supported Klopfer's interpretation. Beck (1945) suggests that the diffuse-shading responses indicate a general withdrawal tendency from the environment and thus, support the shading-passivity hypothesis. He further suggests that *Y* and *YF* responses indicate the experience of an extreme form of withdrawal from the environment leading to complete inability to respond, whereas *FY* responses indicate that the experience is a mild form of withdrawal or passivity leading to a simple delay in responding. There are a number of studies supporting the shading-anxiety hypothesis and the shading-passivity hypothesis, the latter being supported more consistently than the former. Eichler (1951) has found that when the subjects are put under an experimentally induced stress situation, the number of diffuse-shading responses sharply increases. Levitt & Grosz (1960) have demonstrated that hypnotically induced anxiety tends to increase the frequency of the diffuse-shading responses. Other researchers have failed to support the shading-anxiety hypothesis. Holtzman, Iscoe & Calvin (1954) have demonstrated in one study that there is no relationship between the diffuse-shading responses and the scores on the Taylor Manifest Anxiety scale. Goodstein (1954) and Goodstein & Goldberger (1955) reported a similar finding. Elstein (1965) has extended support to the shading-passivity hypothesis, which favours Beck's view. He found a positive relationship between the diffuse-shading responses and the tendency towards withdrawal.

The form dimensional response (*FD*) is relatively a new scoring category introduced by Exner (1974). *FD* responses tend to indicate internalization involving self-inspection or self-awareness. Persons giving such responses are prone to be sensitive and self-critical.

Reflection (*rF* and *Fr*) and pair (*2*) responses were separately scored and analyzed by Exner (1974). His analysis has revealed that reflection and pair responses indicate egocentricity or self-centredness. He has given a formula, which estimates the egocentricity of a person:

$$\text{Egocentricity Index} = 3_r + (2) : R$$

where r includes both rF and Fr and R means the total number of responses on all the ten cards. Hence, egocentricity index $= 3(rF + Fr) + \frac{2}{R}$. The range of the above ratio for a normal adult falls between 0.30 and 0.40. If the ratio falls below 0.30, it indicates a poor self-evaluation due to excessive concern about others and values of the world. If the ratio is above 0.40, it indicates excessive values to the self even at the cost of others and values of the world. In one study of homosexuals, character disordereds and depressives, he demonstrated that homosexuals and character disordereds gave more reflection responses than depressives did. Likewise, homosexuals gave a larger number of pair responses as compared to other groups, whereas depressives gave a fewer number of pair responses as compared to other groups.

Content Scores

Content scores also provide very important interpretive clues regarding the person's needs, interests, interactions and about his various preoccupations. Of the different content scorings, probably the three content scorings, that is, animal, human and anatomy have been widely studied and the evidences regarding them are more precise than evidences on other content categories. Several studies have demonstrated that the animal contents, that is, *A* and *Ad*, represent the single major class of response (Wedemeyer 1954; Neff & Glaser 1954). Animal contents also tend to be influenced by age—children giving the largest number of animal contents and adults giving comparatively a low number of animal contents. In general, an excess of animal contents or higher *A*% (that is, $(A + Ad) \times {}^{100}\!/_{R}$) indicates intellectual constriction and/or emotional disturbances. Normally, *A*% should range from 35% to 45% of the responses. A low *A*% tends to indicate the idiographic character of the person who perceives the objects of the environment in his own peculiar way. Human contents, that is, *H* and *Hd* are the second largest single class of response. Human contents also tend to be influenced by age—children giving fewer human content responses and adolescents as well as adults giving a larger number of human contents. A high frequency of *H* responses indicates better cognitive development and his potential for good relations. A low frequency of *H* responses, on the other hand, is usually associated with delinquency (Ray 1963). Klopfer et al. (1954) argue that the high frequency of *Hd* answers is usually associated with compulsive behaviour and indicates that the subject's preoccupations hinder his interpersonal relationship with others. Some Rorschachers have argued for interpretation in terms of an *H*:*Hd* ratio. For a normal adult, this ratio falls in the range of 2:1 to 3:1. Reversal of this ratio (that is, when *Hd* exceeds *H*) indicates depression, anxiety, constructive defences and intellectual limitations. Content scores like (*A*), (*Ad*), (*H*) and (*Hd*) tend to indicate excessive passivity or withdrawal from the general world of reality into a world of fantasy. In interpretation of animal and human contents, the two ratios are frequently computed and compared:

$$\text{First ratio} = H + Hd : A + Ad$$
$$\text{Second ratio} = H + A : Hd + Ad$$

(These ratios also include (*H*), (*Hd*), (*A*) and (*Ad*) responses.)

Commonly, the first ratio tends to fall in the range of 1:2. Any reversal or imbalance of the ratio tends to indicate impairment of cognitive development and effective social realtions. The second ratio, suggested by Klopfer, gives emphasis to the occurrence of the whole vs detail. It normally falls in the range of 3:1 to 4:1, which indicates ability to perceive things or objects in their right perspective. Reversal in this ratio tends to indicate excessive concern for details often associated with some kinds of pathological traits. The above meaning of the two ratios stands valid and correct only when (*H*), (*Hd*), (*A*) and (*Ad*) do not dominate or occur at great frequency. If such parenthesized responses occur more than the nonparenthesized responses, the ratios become meaningless and parenthesized responses should be treated separately to convey their appropriate meaning.

Anatomy responses also occur frequently in the Rorschach protocol. In general, anatomy responses indicate excess preoccupaton with bodily concerns without any physiological illness. The person releases his affect or motions towards the stress through these preoccupations. It has been reported that psychosomatics usually give more anatomy responses than neurotics (Shatin 1952). A very close associate of the anatomy response is the X-ray response and it also tends to indicate the same form of preoccupation. But the difference is that the preoccupation revealed by the X-ray response is more intense and painful because the reactions of the persons towards the stress are not released, rather kept concealed or suppressed.

Other content scores have been meagrely explained. Sex and blood responses tend to indicate sexual or aggressive acts. Nature, cloud and botany contents indicate emotional deprivation. Responses like fire, landscape, houseold, art, clothing, etc., tend to indicate various kinds of preoccupations causing interference in affective adjustment.

Popular and Original Responses

According to Rorschach, Popular or *P* responses indicate the capability for *conventional thinking*. Beck and Hertz have demonstrated that the high frequency of *P* responses indicates intellectual superiority, whereas the low frequency of *P* responses indicates intellectual retardation. The range of *P* responses for a normal adult lies between 5 to 8. A frequency of *P* responses exceeding the upper limit (that is, more than 8) indicates obsession-compulsion and/or depression. On the other hand, the frequency of *P* responses falling below the lower limit (that is, less than 5) indicates inability to perceive the most common things (which are usually perceived by others), aloofness from the general world, and inability of conformity in thinking. Original or *O* responses may also indicate creativeness or inventiveness of the subject (Exner 1974).

Blend Responses and Organizational Activity

A blend response is one where two or more than two determinants have been used in the formulation of a response. In scoring a blend response, all the determinants constituting the blend are entered and separated by a dot (·), such as $FM \cdot T$, which means that the response contains both the animal movement and the pure texture. Likewise, the blend may include three determinants such as $M \cdot C \cdot V \cdot$, which means that the response contains human movement, pure colour and vista. The majority of the blend responses, however, consist of two determinants. It was Beck who for the first time in 1937 introduced the scoring of the blend responses with the help of a dot (·). Readers should note that the scoring symbols for determinants in the blend answer are the same as would be given if any of the determinants were scored separately. Each determinant in the blends is noted in its original symbol, and determinants which are included in blends are not counted again when the frequencies for each of the single determinant are being counted. The blend response is usually interpreted to indicate complexities in psychological activities at the time of responding. According to Exner (1974), such blend responses indicate a higher level of intelligence. When the frequency of blend responses is very low, it indicates 'undue constriction of the psychological processes'. A blend response is compared with the total number of responses, that is, *R* in the protocol. If the blend response exceeds one fifth of *R*, the frequency of such a response is taken to be high and if the number of blend responses is below one fifth, the frequency is taken to be low.

Organizational activity is another important response on the Rorschach test, though Rorschach himself did not provide a formal scoring for such response. In 1933, Beck became the first person to introduce a formal scoring for organizational activity under the symbol *Z*. The *Z* determinant is the extension of the concept of whole (*W*). It indicates creative ability, that is, ability to perceive and create new wholes. He divided organizational activity into three subcategories, depending upon the type of organization and the complexity of the stimuli, each subtype being weighted differently. The symbol for all types of organizational activity was, however, the same *Z*. The three types of organizational activity are as under:

1. An organizational activity which meaningfully integrates two or more adjacent detail areas
2. An organizational activity which meaningfully integrates two or more nonadjacent detail areas
3. An organizational activity in which white space is meaningfully integrated with details of the blot

One important step to be taken in scoring a response as Z is that form must always be involved either in a primary or secondary way. This automatically reveals that pure colour response (both chromatic and achromatic), pure texture response, pure vista response and pure diffuse-shading response would never be scored as Z. In 1940 Hertz also introduced scoring for organizational activity in her system but the different types of organizational activity, unlike Beck, were weighted equally. Organizational activities are interpreted as indicating intellectual activity. Hertz, however, interpreted organizational activities as the composite representative of *intellect, creativeness, drive* and *efficiency*. A very low frequency of Z scores indicates depression and anxiety in persons who are of superior intelligence.

For making a full and coherent interpretation of a Rorschach protocol, it is also essential that the five main variables introduced by Rorschachers be clearly understood. The five variables are: the Erlebnistypus (*EB*), the Experience Actual (*EA*), the Experience Potential (*ep*), the Experience Base (*eb*) and the Blend response. The last variable has already been explained and interpreted. Therefore, in the following paragraphs only the first four are discussed.

The Erlebnistypus or Erlebnistyp (*EB*) is one of the important ratios suggested by Rorschach in 1921. It is also referred to as the Experience Type or the Experience Balance or movement–colour sum ratio or $M{:}\Sigma C$. The *EB* respresents a comparison of the sum of human movement determinant (M) to the weighted sum of each chromatic colour determinant (that is, C, FC, CF and Cn). The sum weighted C is calculated by assigning the following values to each colour determinant, $FC = 0.5; CF = 1.0$; C or $Cn = 1.5$. It may also be written as:

$$\Sigma C = \frac{FC + 2CF + 3C}{2}$$

Rorschach's contention, which is still supported by many studies, denotes that the movement–colour sum ratio or the *EB* indicates basically two types of constitutionally predisposed response tendencies. If the M of $M{:}\Sigma C$ is greater, it indicates *introversiveness* or proneness to use one's inner life for the satisfaction of basic needs. On the other hand, if the ΣC of $M{:}\Sigma C$ is greater, it indicates *extroversiveness* or proneness to use social interactions as a means for gratification of one's own important needs. Such a person likes more to express his affect towards the world than an introversive person. Some Rorschachers have prefixed a certain ratio and when M outweighs that ratio, introversiveness is indicated, and when ΣC outweighs that ratio, extroversiveness is indicated. For example, Klopfer et al. (1954) have said "when M responses outweigh the ΣC by 2:1 there is an introversive balance; whereas when ΣC outweighs by 2:1 there is an extroversive balance." Rorschach described a rare possibility in which both sides of the ratio are equal, that is, M becomes equal to ΣC. He called such persons *ambiequal*. Such persons share the experiences of introversiveness as well as extroversiveness, and therefore, they are described as showing failure to establish a certain response tendency for the gratification of basic needs.

The Experience Actual *(EA)* is the second important variable of the interpretative significance. The *EA* was for the first time introduced by Beck (1960). The *EA* is obtained from the data shown in the *EB* and is equal to the sum of all components appearing in *EB*. It may be written as under:

$$EA = M + \Sigma C$$

Beck has pointed out that *EA* indicates that the affective experiences of persons are organized or controlled.

The Experience Base (*eb*) represents a ratio which is based upon a similar ratio suggested by Klopfer et al. (1954). The *eb*, also known as the confirmatory ratio, is nothing but a comparison of all nonhuman movement determinants with all categories of pure shading determinants and achromatic determinants. Thus *eb** may be written as indicated below:

$$eb = \Sigma(FM + m):(T + V + Y + C')$$

According to Klopfer, the *eb* indicates a kind of response tendency (either introversive or extroversive), which has not been fully accepted by the persons at a given time. The introversive tendency is reflected by $(FM + m)$ and extroversive tendency is reflected by $(T + V + Y + C')$. Both of these response tendencies are such that they are not easily controlled by higher cognitive actions. It is obvious from the above facts that the interpretive significance of *EB* and that of *eb* are similar to some extent. Ordinarily, *eb* should follow the same direction as *EB* because it also indicates the above two types of response tendencies. If this happens, the implications of *eb* are said to have been strengthened and confirmed.

The Experience Potential (*ep*) is another important variable of interpretative significance. This ratio is based upon derivation of data incorporated in *eb* and is equal to the sum of all nonhuman movement determinants plus all varieties of pure shading determinants and achromatic colour determinants. Thus *ep** may be written as under:

$$ep = \Sigma(FM + m) + \Sigma(T + V + Y + C')$$

Thus the *ep* (Experience Potential) is related to the *eb* (Experience Base) in the same fashion as the *EA* (Erlebnistypus) is related to the *EB* (Experience Balance). The *ep* tends to indicate such needs and affects which are not properly organized or controlled.

As students may often get confused when reading the discussion on the interpretation of the Rorschach responses, the author has provided a synthesized general outline to help them in interpreting scores (see box on next page).

There is little agreement regarding the scientific status of the Rorschach tests. The opponents claim that the test lacks a universally accepted standard of administration, scoring and interpretation; evaluation of data are subjective; results are unstable over time; it is unscientific and is inadequate by all traditional standards (Exner 1995, Nezworski & Wood 1995, Wood, Nezworski & Stejskal 1996).

The reliabilities and validities of the test have also been debatable. Commenting upon the specific level of validity of the test, Buros (1970) has very clearly said, "The vast amount of writing and research has produced astonishingly little, if any, agreement among psychologists regarding the specific validities of the Rorschach." Meta-analysis by Parker (1983) reported an overall internal reliability coefficient of .83 which is based upon 530 statistics from 39 papers published in between 1971 and 1980. However, meta-analysis by Parker was criticised on the grounds that results on validity were not analyzed separately from results on reliability (Garb, Florio & Grove 1998).

Lastly, taking the pros and cons, it can be safely concluded that a final word on the Rorschach test is yet to come. More researches are needed but unless the practitioners can agree on a standard method of administration and scoring, the hands of the researchers would remain tied.

---

* For Klopfer, $\Sigma(T + V + Y + C')$ is substituted by $\Sigma(FC + C + C')$ in which *FC* and *C* represent texture responses and *C'* represents achoromatic colour responses.

A general outline for Rorschach interpretation

Total responses (R) =

Total time (T) =

Average time per response (T/R)=

Average reaction time for achromatic series =
(Cards I, IV, V, VI and VII)

Average reaction time for chromatic series =
(Cards II, III, VIII, IX and X)

$$\text{Affective ratio} = \frac{\text{number of responses to cards VIII, IX and X}}{\text{number of responses to the remaining 7 cards}}$$

$$\text{Egocentricity index} = \frac{3(Fr + rF) + (2)}{R}; F\% = \frac{\Sigma F}{R} \times 100$$

$$\text{Lambda Index} = \frac{\Sigma F}{\text{sum of non}-\text{pure } F} =$$

$$F^{+}\% = \frac{\Sigma F^{+}}{(\Sigma F^{+}) + (\Sigma F^{-})} \times 100 =$$

$$FC : CF + C =$$

$$A\% = \frac{A + Ad}{R} \times 100 =$$

$$H\% = H + \frac{H + Hd}{R} \times 100$$

Blends : R =

$W : M$ =

$W : D$ =

Number of P =

Number of O =

Frequency of Z (or Z_f) =

$(H + Hd) : (A + Ad)$ =

$(H + A) : (Hd + Ad)$ =

$$\Sigma C = \frac{FC + 2CF + 3C}{2} =$$

Erlebnistypus, or $EB = M : \Sigma C =$

Experience actual, or $EA = M + \Sigma C =$

Experience base, or $eb = \Sigma(FM + m) : \Sigma(T + V + Y + C') =$

Experience potential, or $ep = \Sigma(FM + m) + (T + V + Y + C') =$

THE HOLTZMAN INKBLOT TEST

Modelled after the Rorschach test, Holtzman et al. (1961) developed an inkblot test which is known as the Holtzman Inkblot Test (HIT). Holtzman developed this test in order to remove some of the basic technical difficulties of the Rorschach test like unlimited number of responses, poor scorer reliablity, etc. The test is available in parallel forms, each form (Form A and B) having 45 cards. Both coloured and achromatic cards are included and a few inkblots are markedly asymmetric. The subject is permitted to give one response per card and thus, the total number of responses for any examinee is a fixed one, thereby automatically avoiding many of the difficulties in scoring of the Rorschach test. Each response is followed by a two-fold simple question: Where was the percept represented in the blot and what the percept suggest about the blot? All the responses are classified under 22 response variables, which include the Rorschach variables plus variables like hostility, anxiety, and pathological verbalization. The main advantages of the Holtzman test over the Rorschach test are objective and easy intercomparability of examinees because of the limited and fixed number of responses and availability of percentile norms for all the 22 variables for samples ranging from 5-year-olds to adults. Due to the above factors, many researchers have, in general, agreed that the Holtzman test appears to be better standardized than the Rorschach test. A group-test form of the Holtzman inkblot test to be administered to more than one individual at a time has also been recently developed (Holtzman et al. 1963).

The major HIT scoring variables can be discussed as under:

1. *Reaction time*: It includes time in seconds beginning from the presentation of the inkblot to the beginning of the main response.
2. *Rejection*: It is scored when the examinee or testee fails to report anything or returns the inkblot to the examiner.
3. *Location*: It is scored on a 3-point system: 0, 1 and 2. 0 is scored for whole blot, 1 for large area and 2 for smaller area.
4. *Space*: It is scored when there is a true figure-ground reversal where white part is the figure.
5. *Form Appropriateness*: It is scored on three points: 0, 1 and 2; 0 for poor, 1 for fair and 2 for good. It is a sort of goodness of fit of the concept to the form of the blot.
6. *Form Definiteness*: It is scored on 5-point scale ranging from 0 to 4; 0 for formless concept and 4 for highly formed concept
7. *Shading*: It is scored 0 to 2. Shading such as texture, fuzziness are scored as determinant
8. *Colour*: It is a primary determinant and scored 0 to 3.
9. *Movement*: It is scored 0 to 4 and is reserved for those responses which implies energy or dynamic movement quality.
10. *Pathognomic Verbalization*: It is scored for absurd, queer and incoherent verbalizations to cards.
11. *Integration*: It is scored 0 to 1. When two or more blot elements are effectively integrated in the response, it is scored 1 and otherwise it is scored 0.
12. *Content scores*: It includes categories like Human, Animal, Anatomy, Sex, Abstract and is scored 0 to 2 based upon absence, partial or full presence of the concept.
13. *Anxiety*: Each response is scored 0 to 2 for signs of anxiety.
14. *Hostility*: Each response is scored 0 to 3 for showing the signs of hostility.
15. *Barrier*: It refers to any protecting covering that might be symbolically related to body image boundaries. If barrier is present, it is scored as 1 and if absent, it is scored as 0.
16. *Penetration:* It is scored 1 if the concept is symbolic of the testee's feeling that his or her body's external parts can easily be penetrated. In otherwise case, it is scored 0.
17. *Balance:* It is scored 1 if the testee sees presence or absence of symmetry in the design. Otherwise, it is scored as 0.
18. *Popular:* It refers to commonness of the response and scored 1 if the response is common. Commonness is determined on the basis of 1 in 7 normative protocols.

Administration and Scoring of HIT are well standardized. Its scorer reliability appears to be highly satisfactory. The various other reliabilities like split-half, alternate-form and test-retest reliability investigations although show some differences, but their results are encouraging. Validity data on HIT also have yielded satisfactory results.

A recent variant of the HIT called the HIT 25, a short version, consisting of the first 25 cards from Form A of HIT with two responses per card has been developed by Holtzman (1988). This shorter version, that is, HIT 25 has been found very useful for diagnosing the cases of schizophrenia. A group form of the HIT, using slides, has also yielded scores on most variables that are well comparable to those obtained through individual administration (Holtzman, Moseley, Rienehr & Abbott 1963; Swartz & Holtzman 1963).

Cassell's Somatic Inkblot Series (SIS)

Cassell's SIS (1965, 1979, 1980, 1984) is another important latest development in the field of inkblot tests. He was not satisfied with the stimulus materials of earlier inkblot tests through which the sufferings of the patient were studied. These earlier tests were useful for understanding only the basic structure of personality. He appreciated the idea that emotional complaints of the patients could be easily understood through somatic processes.

SIS, according to Cassell, is a sort of structured projective diagnostic test. It is projective because it is based on spontaneous, individualized responses to semi-ambiguous inkblot figures. It is structured because of its sequential presentation of research and with typical or atypical response potential. It is diagnostic test in the sense that it makes interaction of structure and stimuli that evokes meaning and symbolism unique to individual and this helps in differentiating typical and atypical persons.

There are three forms of SIS :

(a) *Somatic Inkblot Series I (SIS-I):* It consists of 20 inkblot images printed on cards like Rorschach test.

(b) *Somatic Inkblot Series II (SIS-II):* It consists of a 8-page booklet with 62 images printed on it. Each image is provided with a space below it for writing the response.

(c) *SIS-Video:* This is really a video version of SIS-II where video starts with pictures of flowers and soft music followed by images one by one. Each image stays on the screen for about 30 seconds. It is also designed for self-administration where the examinee writes answers on a sheet.

Cassell's explanations were further crystallized and reinforced by Piotrowski (1957), Lerner (1991) and Schafer (1954).

THEMATIC APPERCEPTION TEST (TAT)

The Thematic Apperception Test, also known as TAT, is another projective test commonly used in clinical and nonclinical settings. The TAT was first published by Murray in 1935, under the head 'A method for investigating fantasies: the Thematic Apperception Test' in *Archives of Neurology and Psychiatry.* Later on, Murray & Morgan (1938) working at the Harvard Psychological Clinic published a book, *Exploration in Personality,* in which the details of analysis of TAT appeared. As compared to the Rorschach test, the TAT has less vague and ambiguous pictures to which the examinee responds. As the RT and the TAT provide complementary information, a combination of these two has proved to be very effective, though the latter has been found to be much more effective in the comprehensive study of personality and in the interpretation of neurosis, psychosis, behaviour disorders and psychosomatic diseases. According to Murray (1971, 1) the purpose of the TAT is to reveal "some of the dominant drives, emotions, sentiments, complexes, and conflicts of personality. Special value resides in its power to expose the underlying inhibited tendencies which the subject or the patient, is not willing to admit or cannot admit because he is unconscious of them."

In the TAT, two terms are worth mentioning, namely, thematic and apperception. The term 'thematic' has been derived from the term 'thema' which refers to a subject or topic on which a person thinks, speaks or writes. Murray has defined thema in a much broader sense which, of course, includes the above meaning. According to him, thema is defined as an interaction of the need and press variable. Murray has defined need as a hypothetical process within the organism, which stimulates him into either covert or overt action. Similarly, Murray (1971) has defined press as a force in the environment, which may facilitate or interfere in the satisfaction of the need of the organism. The concept of thema is, thus, based upon this 'need–press' theory and simply represents a patterning of these two basic variables, that is, need and press. The second term of the test 'apperception', refers to a clear perception involving definite recognition or indentification. Thus, apperception is different from perception in the sense that the latter may sometimes be vague or indistinct. The inclusion of the term makes justification for the fact that thema is apperceived, that is, the examinees not only perceive but also recognize the implications of the stimulus situation of the card. In this sense, the TAT is different from the Rorschach test because in the latter, examinees only perceive the stimulus of the card without making any definite recognition of its implications.

The standard TAT consists of 19 cards containing vague or ambiguous pictures in black and white plus one blank card. There are thirty-one cards (30 pictured cards plus one blank card) in a series and these cards are used in various combinations depending upon sex and age. The maximum number of cards to be administered to any one individual is 20. Murray recommends that TAT has no utility for children below four years of age. Some cards are used with all subjects and some are used with one sex group or with a particular age group. Cards for a specific sex group or age group are distinguished by letter symbols: F for adult females, M for adult males, B for boys under 14, and G for girls under 14, BM for both boys under 14 as well as adult males, GF for both girls under 14 as well as adult females, BG for both boys and girls under 14. Cards having none of these symbols such as 1, 2, 4, 5, 14, 15, etc., are meant for both sexes and all ages. All the 20 cards are usually administered in two one-hour sessions, 10 cards in each session. The more unusual, dramatic and bizarre type of cards are usually selected for the second session, that is, pictures selected for the second session are less like everyday life situations. Murray (1971, 5) recommends an interval of at least 24 hours between the two sessions. Cards to be used in the first session are 1, 2, 3BM, 3FG, 4, 5, 6BM, 6GF, 7BM, 7GF, 8BM, 8GF, 9BM, 9GF, and 10 whereas the cards to be used in the second session are 11, 12*M*, 12*F*, 12*BG*, 13*MF*, 13*B*, 13*G*, 14, 15, 16 (blank card), 17*BM*, 17*GF*, 18*BM*, 18*GF*, 19 and 20. In a clinical setting, most clinicians use only 10 cards selected according to the purposes at hand. Murray (1971) has provided for separate instruction to normal adults and children under 14. These instructions are approximately given as follows:

Instruction for Normal Adults in the First Session

> This is a test of imagination. I shall show you some pictures, one at a time; and your task will be to describe a story. Your story must include what has led up to the event shown in the picture, what is happening at the moment, what the characters are feeling and thinking; and the outcome. Express your thoughts as they come to your mind. Do you understand? You can devote about five minutes to each story. Here is the first picture.

Instruction for Children under 14, Or Poorly-educated Adults, Or Psychotics in the First Session

> This is a story-telling test. I have some pictures, which will be shown to you and for each picture I want you to tell a story. Your story should include what has happened before and what is happening now. Also write what the people are feeling and thinking and how it will come out. You can write any kind of story you want. Do you understand? Here is the first picture. You have five minutes for writing a story. See how well you can do it.

Instruction for Normal Adults in the Second Session

> The procedure today is the same as before. Your first ten stories were excellent and coherent but you mostly kept yourself confined to the facts of everyday life. Now I would very much appreciate seeing what you can do when you disregard the realities of everyday life and let your imagination have its expression, as in a myth, fairy story or allegory. Here is picture No 1.

Instruction for Children under 14, Or Poorly-educated Adults, Or Psychotics in the Second Session

> Today I shall show you some more pictures. These pictures are much better and more interesting. You told me some fine stories the other day. Now I want to see whether or not you can describe a few more. Describe them in an even more exciting way than you did last time if you can—like a dream or fairy tale. Here is the first picture.

Card No. 16, which is a blank card, is preceded by a special instruction like this:

> See what you can see on this blank card. Imagine some picture there and describe the story.

The responses of the examinee are recorded verbatim and comments by the examiner are allowed after the story for the first card is completed. The purpose of comments is to know whether or not an outcome has been mentioned, whether the examinee has taken a very long or short time in completion of the story and whether the story is very long or short. (The average length of a TAT story is 300 words. A story having less than 140 words indicates lack of co-operation and involvement by the examinee and hence is not worth scoring.) After this first story, Murray (1971, 4) recommends, "...as a rule it is better to say nothing for the rest of the hour." However, the examiner is left free to intervene in the remaining stories if he thinks that intervention would improve the specific content of the story without actually influencing the content of the story itself. The intervention becomes absolutely necessary when the examinee is taking a longer time than necessary or a very short time in describing the story and/or if he leaves out a very important part such as the outcome of the story. If the examinee describes several stories rather than one story in a single picture, he should be encouraged to concentrate on only one long story rather than several stories. A too long story may be cut off by the examiner (Murray 1971, 4). In this way all the ten cards are given successfully and thus, the first session is over. After a time gap of one day, or 24 hours, the second session starts, which is allowed to be completed in the same way. The cards of this session resemble less with everyday life situations and an information to this effect is given to the examinee. In this session, the blank card is given to the examinee and here the examiner gives a slightly modified version of the instruction. When all the 20 cards have been given, Murray (1971) recommends a process of interview which may be held either immediately or after a few days. The purpose of the interview is to know what sources or associations the examinee has used for the stories—whether the source of the ideas of the stories is his private experience, or novels or movies or the experiences of relatives and friends.

The scoring and interpretation of the TAT stories is the next important task to be accomplished by the examiner. The scoring is not separable from the interpretation in the TAT, and thus, it is different from the Rorschach test where it is. Not only this, in scoring the Rorschach responses, most Rorschachers more or less agree with regard to the symbols and the criteria used, whereas in the TAT no such agreement exists among the TAT users. As a matter of fact, no formal scoring (except Hero, Need, Press, Thema and Outcomes) of any fixed kind is used by the TAT users (Nunnally 1970). In such a situation, Murray's own proposals regarding the interpretation of the TAT seem most appropriate. The TAT differs from the Rorschach test in one other basic sense, too. The purpose of the Rorschach test is to reveal the structure and organization of personality, whereas the purpose of the TAT is to reveal the contents of the personality such as drive, need, fantasies, complexes, conflicts. The TAT is based upon the assumption that when the examinee interprets the ambiguous scenes by way of creating stories, he organizes the various contents of his personal experiences, which reveal the different aspects of his own personality.

In interpreting the TAT stories, Murray makes the following two basic assumptions:

1. The characteristics of the hero (also called the *focal figure* by some TAT users) of the story broadly represent the tendencies of the examinee's personality.
2. The situation which surrounds the hero of the story represents the various facets of the examinee's life situation containing his past, present and future environment.

These two assumptions are analyzed under the following six categories:

The Hero

The hero (either male or female) of the story is the central character about whom the entire events revolve. The interpreter recognizes the hero of the story by his principal traits like superiority, inferiority, leadership, belongingness, mental abnormality, criminality, quarrelsomeness, etc. It is the hero of the story with whom the examinee identifies himself and in whom he is most interested. The hero of the story plays the leading role in the story and is the main character who is associated from the beginning of the events to the outcome of the story. A story may have one hero or sometimes two, three or four heroes, which Murray has called a *sequence of heroes*. Sometimes, the hero characteristics are equally divided among equally significant characters. Such heroes are known as partial heroes. Sometimes, the story is such where the hero only listens to the events in which another character is involved. Such stories, therefore, contain a primary hero (who hears the events) and a secondary hero who narrates the events.

Motives, Trends and Feelings of the Heroes

Under this head, Murray explains that the interpreter should carefully analyze overt behaviour (both usual and unusual types) of the hero (that is, what the hero thinks, feels and does) because these overt behaviours reflect the needs (or drives) and emotions of the hero (and, therefore, of the examinee). A need may be expressed subjectively in the form of a wish, intention or impulse of the hero or it may be expressed objectively in the form of a trend of overt behaviour. Murray has given a list of 28 needs. Some of these are: *n* aggression divided into four parts—emotional and verbal, physical and social, physical and asocial, and destruction; *n* achievement, *n* passivity, *n* dominance, *n* abasement, *n* nurturance, *n* sex, *n* succorance, and *n* aggression, *n* acquisition, *n* affiliation, *n* autonomy, *n* deference, *n* cognizance, *n* blame avoidance, *n* harm avoidance, etc. The letter *n* is the abbreviation used for need. Murray further suggests that the manifestation of these needs should be rated on a 5-point scale ranging from 1, representing a mild form of the need, to 5, representing the intense form of the need. In recognizing the various needs, due attention is also given to the intensity, duration, frequencies of their occurrence in different stories and the general significance of the story. After all the 20 stories are scored on a 5-point scale, a total score for each need is computed and compared with the standard score for the examinee of given age and sex. Those needs whose total scores are very high or very low are carefully analyzed in relation to each other. Some of the interpretive hints are like this: A male examinee showing unfriendly motives and actions to female characters in most of the TAT stories indicates that he has some troubled relations with women in his life. Likewise, if a hero of most of the stories displays disappointment, embarrassment and failure, it shows that the person feels defeated and depressed in his actual life. Special attention is also paid to the unique materials, which significantly differ from the common and popular responses to each picture because these unique and unusual materials may have special significance and relevance in the examinee's life.

Forces of the Hero's Environment

Under this head of the TAT stories emphasis is placed upon environmental variables or forces (called press), which can either facilitate or interfere with the gratification of needs of the hero. These environmental variables include details of objects and persons noting their uniqueness, intensity and frequency. Special attention is also paid here to include objects and persons which are not in the picture but invented by the examinee. More than 30 such press (plural is also press) have been listed by Murray. A few examples are *p* rejection, *p* dominance (divided into three parts—coercion, restraint, inducement or reduction), *p* physical injury, *p* lack or lost, *p* aggression (divided into four parts—emotional and verbal, physical and social, physical and asocial and destruction of property), *p* affiliation (divided into two parts—associative and emotional), *p* nurturance, *p* physical danger (divided into two parts—active and in support), etc. (The letter *p* is based on the analogy *n* and refers to press.) All these environmental forces (or press) are classified according to their effect (promised or realized) upon the hero. The strength of press in each story is rated on a 5-point scale of which 1 indicates the minimum strength and 5 indicates the maximum strength. Like the criteria of need, here also the criteria are intensity, duration, frequency and general significance in the story. After rating all the 20 stories on the 5-point scale, the total score for each press is computed and then, it is compared with the standard scores for the examinee of a given sex and age. The total scores, which are very high or low as compared with standard scores, are carefully examined in relation to each other.

Outcomes

Outcomes refer to how the story concludes. There may be a happy ending or an unhappy ending. In assessing outcomes, the comparative strength of the forces emanating from the hero and the strength of the forces emanating from the environment are compared and analyzed; the magnitude of frustration and hardships shown by the hero as well as his success or failure in handling those situations is also analyzed.

Themas

Themas refer to the interaction of need (s) of the hero with environmental forces, that is, press, combined with the successful or unsuccessful outcome arrived at by the hero of the story. The purpose of a thema is to study the different types of forces in their relationships and to determine the most important problem of the examinee from the interaction of needs and factors leading to the gratification of those needs. Obviously, then, a thema is nothing but a synthesis of factors analyzed into hero, need, press and outcome. Themas may be simple or complex.

Interests and Sentiments

Under this heading, choice of topics of the story and ways of handling them are analyzed. The fundamental assumption here is that the hero's interests and sentiments as revealed by the choice of the topics and the way in which they are handled are the examinee's own interests and sentiments.

Other TAT interpreters like Tomkins, Wytt, Rapaport, Bellak, Arnold, etc., have suggested different techniques for analyzing the stories. Among all the different techniques, Murray's technique with some modifications is popular and frequently used. The most widely used categories for analyzing the TAT stories are hero, need, press, outcomes and themas.

In arriving at a conclusion on the basis of the analysis of TAT stories, the following points should be kept in view:

1. Normally, about six out of the twenty stories, that is, approximately 30% of the stories should fall under the impersonal category. This category consists of impersonal elements like the elements given in the picture, elements invented at the moment, fragments read in novels or witnessed in movies, etc., none of which represents the important determining tendency of the personality of the examinee. If more than 30% of the stories fall in this category, it indicates that either the examinee is not involved in the task or the content is psychologically irrelevant or the test itself has not been skillfully administered.

2. Murray (1971) distinguishes between two levels of functioning: the first level which includes the physical and overt verbal behaviour, and the second level which includes plans,

ideas, dreams, fantasies, etc. TAT has been designed to reveal the second level and not the first level of functioning, although on the basis of the second level a guess regarding the first level can be made.

3. Murray (1971) has distinguished between three layers of normal personality—outer layer, middle layer and inner layer. The inner layer consists of repressed tendencies or wishes, which are never (or rarely) expressed in thought (that is, the second level), but expressed in action (the first level). The middle layer consists of desires and tendencies, which can be expressed in thought in undisguised form (the second level) and can sometimes, also be expressed in action (the first level). The outer layer consists of those wishes, desires and tendencies, which are frequently expressed publicly (the second level) and openly expressed in behaviour (the first level). The skillful interpreter must find out to which of these layers the different important variables such as need, press, etc., belong. Generally, the stories described in response to the first ten cards (given in the first session) tend to reveal the outer layer of the personality and the stories described in the second session regarding the remaining ten pictures tend to express the inner and middle layer of personality.

The above points should be carefully dealt with in arriving at a final conclusion. Murray warns that any conclusion arrived at on the basis of TAT stories should be regarded only as working hypothesis (to be verified by other techniques) rather than proved facts. As Murray (1971, 15) has said:

"The TAT draws forth no more than twenty small samples of the subject's thoughts. To suppose that these will invariably provide a skeleton of the total personality is unduly optimistic...there are sets of TAT stories...from which it is impossible to infer the underlying determinants of character."

In India, TAT has been adapted by Dr Uma Chaudhary (1960). This adapted version of TAT has 14 cards which have environmental details and human figures in context of Indian set up. Out of these 14 cards, card No. III and Card No. IV have got gender reservation. For example, Card No. III M.B. is reserved for male testee whereas Card No. III FG is reserved for female testee. Likewise, Card No. IV has also MB and FG series. Other Cards are common for both genders.

Derivatives of TAT

Like the Rorschach test, there are some derivatives of the TAT. Modelled after the TAT, a few pictorial tests have been developed. The Children's Apperception Test (CAT) is one such test. The CAT, developed by Bellak (1954), is meant for children between three and ten years. The test consists of ten cards, which are slightly smaller than the standard TAT cards. The CAT pictures substitute animals for human beings, the assumption being that animals tend to produce fantasies more readily than human beings among children and therefore, children can easily project to animal pictures as compared to pictures of human beings. These animal pictures are, however, presented in anthropomorphic fashion. All the CAT pictures intend to arouse fantasies relating to sibling rivalry, toilet-training, aggression, etc. Recently, Bellak & Hurvich (1966) have prepared a human modification of the CAT, which is meant for use with children beyond 10 years. Older children are more capable of projecting to human figures and as such, the human modification of the CAT (known as CAT-H) was prepared. The Rosenzweig Picture-Frustration (P-F) Test is another test modelled after the TAT. The P-F test, developed by Rosenzweig in 1942, is slightly different from the TAT and the CAT. The latter are picture-story tests whereas the P-F test is a test where the examinee is required to write a short bit of conversation in the caption box of each picture. The test has two forms—adult's form (for 14 years and older) and children's form (for 4 to 13-year-old children). In both the forms each picture consists of cartoon-like drawings having two principal characters. One of the principal characters is engaged in a frustrating situation common in life and the other character is saying something, which calls attention to that frustrating situation. The examinee is requested to write in the caption box of each picture what the reply would be of the frustrated character. He is also instructed to write the very first reply that comes to his mind. In administering the P-F study, it is assumed that the examinee identifies himself with the frustrated character and projects his own frustrated tendencies in the form of the reply of that character. The P-F test is based upon the frustration–aggression hypothesis, which stipulates that frustration always leads to aggression but since the individual has the capacity for frustration tolerance, the aggression produced by frustration may not be expressed outwards or may even be derived. The frustrating situations included in the P-F study are of two types—ego-blocking and superego-blocking. An ego-blocking situation is one in which some obstruction, mostly personal in nature, disappoints one of the two principal characters directly, and a superego-blocking situation is one where one of the two principal characters is directly insulted or incriminated by the other characters.

The P-F study aims at studying the types of aggression and the direction of aggression. According to Rosenzweig, there are three types of aggression and three types of direction of the aggression. The three types of direction are: *extrapunitive* or directed outwards or towards the environment (also known as extra-aggression); *intropunitive* or directed inwards or towards the self (also known as intra-aggression); and the *impunitive* or evading the frustrating situation or denying its presence. The three types of aggression are: obstacle-dominance in which the emphasis is upon basic events which acted as an obstacle in producing the frustration; ego-defence in which the emphasis is upon the protection of experience of the frustrated individual; and need-persistence in which the emphasis is upon solution or remedy of the problem which produced the frustration. In scoring the responses of the P-F study, the percentages of the responses falling into each of these three categories of direction and types of aggression are computed, and then compared with some normative percentages.

Other TAT derivatives include the Object Relation Technique, the Pickford Projective Pictures (PPP), the Blacky pictures, etc. A detailed discussion of these varieties is beyond the scope of this book.

VERBAL TECHNIQUES

Verbal techniques are those techniques in which stimulus materials are verbal (and not pictorial), and where the examinees are also requried to give their responses verbally towards those stimulus materials. Verbal techniques are different from pictorial techniques in the sense that in the former, both the stimulus and the response are verbal, whereas in the latter the response is verbal and the stimulus is pictorial. The word-association test and the sentence-completion test are the best examples of verbal techniques.

Word-Association Test

The word-association test is one of the most popular projective devices for studying personality. This test requires the examinee to tell the very first word that comes to his mind after listening to the stimulus word by the examiner. The examiner notes down the responses and the reaction time taken to respond towards each stimulus word.

The word-association test was first described by Galton in 1879. Later on, the early experimental psychologists like Wundt and Cattel introduced this technique to study thinking processes. The clinical psychologists also became interested in the technique and started using it as a means of exploring emotions, conflicts and anxieties. Kraepelin used this technique in the exploration of personality of mentally ill persons. However, Jung in 1910 made a systematic effort to introduce the technique in his psychoanalysis. He presented a standard list of 100 words (some neutral and some emotionally toned) and analyzed the responses of the examinees from different angles to get a variety of diagnostic clues. His main analysis included the analysis of time taken in responding (reaction-time) and the content of the responses. A longer time was taken to be indicative of some conflicts, anxieties, repressed experiences and a sign of emotional embarrassment. Delays indicated that those words touched off painful experiences and threatened to bring to light the anxiety-provoking or guilt-laden materials or were close to the repressed experiences. The content of the responses is analyzed from the angle of contrast (black-white), being unusual (chair-sea), supraordinate (cow-animal), clang association (land-hand), etc. A large number of unusual responses such as 'Pen-Bear', etc., are taken to be indicative of mental illness. Jung had also studied the retest behaviour in the word-association test. He obtained the responses towards his preselected words (which included both neutral and emotional words) and then, the examinee was again administered the test with the instruction to recall the responses given previously. Changes in responses were taken as important clues for further exploration of emotional complexes of the individual. Later on, Kent and Rosanoff also developed their own test for use as a screening instrument in psychiatric clinics. The test is known as the Kent-Rosanoff Free Association Test, which consists of 100 words belonging to the neutral and common category. On the basis of the responses to each word of the examinees, the frequency tables of the responses of each word are prepared. An 'Index of commonality' (or the median frequency value) is calculated on the basis of the frequency tables. Kent and Rosanoff found that mentally ill persons exhibit a lower index of commonality, that is, they tend to give uncommon or individual responses (or less common responses) more than the normal ones.

Still another important word-association test was developed by Rapaport and his associates in 1946. The test consists of 60 words and aimed at exploring the areas of aggression and sexuality with special emphasis upon oral, anal and phallic levels of psychosexual development.

Sentence-Completion Test

The sentence-completion test consists of a series of partial or incomplete sentences and the examinees are required to complete them. The examinees are free to take a decision in completion of the meaning, and no time limitation is imposed for their decision. Originally, sentence completion was used by Ebbinghaus as a nonprojective technique of assessing intellectual level. A serious attempt to use this test to assess personality traits began in the 1930s, and one of the earliest tests of this type was developed by Rohde & Hidreth in 1940. Subsequently, several sentence-completion tests were developed. A few examples of sentence-completion items are as follows:

I feel tense ....
I feel that sex ....
My sex life ....
I feel guilty about ....
Sex relations ....

The responses on the sentence-completion tests are analyzed in a more or less similar way to those of TAT stories, that is, in analyzing these responses, the motives, expectations, moods, etc., are interpreted in order to arrive at a meaningful conculsion. A few popular sentence-completion tests are the Rotter Incomplete Sentences Blank, OSS test, etc.

EXPRESSIVE TECHNIQUES

Expressive techniques are those where the examinee is given an opportunity to express his personality in the form of certain drawings, finger-paintings, play, role play, handwriting, and free use of toys, etc. In such techniques the ways in which the individual manipulates or works with the given materials are important (and not the end product of these manipulative acts) because it is assumed that through these manipulations he expresses his needs, motives, emotions, conflicts, etc. Some examples of expressive techniques are the Machover Draw-a-Person (D-A-P) test, Buck's House-Tree-Person (H-T-P) test and the World test, etc. Some of the important expressive techniques may be studied under the following headings.

Figure-Drawing Tests

In figure-drawing tests, the examinee is given a sheet of paper and pencil to draw the figures of a person and in such drawings he is assured by the examiner that this is no longer a test of his drawing ability and thus, is encouraged in his efforts. The two most common figure-drawing tests are the Machover Draw-a-Person (D-A-P) test and Buck House-Tree-Person (H-T-P) test. In the D-A-P test, the examinee is instructed to draw a person. If the first drawing is male, he is asked to draw a female and vice versa. The examiner systematically notes down the sequence in which the different parts of the body are shown. He also notes down the other comments and procedural details. The drawings are sometimes followed by inquiry to elicit specific information regarding the details of drawings. In an H-T-P test the examinee is required to draw the figure of a house, a tree and a person. According to Buck, the figure of the house is taken to be the indicative of the associations concerning the examinee's home and those living therein; the figure of the tree, of the associations relating to his life role and his ability to derive gratification from the general environment; and the figure of a person, of the interpersonal relationship of the examinee.

All such figure-drawings (emerging from the D-A-P test and H-T-P test) are interpreted from three angles: (i) analysis of the general overall impression of the drawings; (ii) analysis of the structural features of the drawings; and (iii) content analysis of the drawings. The analysis of the general overall impression of the figure concerns itself with the posture of the figure (giving the impression of action or being static), the facial expression, and other similar features. The overall general impression may reveal the impression of expansiveness, hostility, aggressiveness, and submissiveness. The structural analysis of the figure is concerned with factors like the size of the figures, the presssure of the lines, the sequence of the parts drawn, their position on the given page and the like. The size of the figure is said to be directly related to the self-esteem of the examinee. The examinee who thinks about himself in a degrading fashion (poor self-esteem) is likely to draw a very small figure whereas the examinee who thinks about himself with an air of superiority, often draws a big figure, sometimes so big that it requires two sheets of paper for a single figure. The content analysis of the figures is concerned with emphasis upon the different parts of the body given in drawings and depiction of clothing and accessories. It has been demonstrated that emphasis upon drawings of head at the expense of other parts of the body reveals several important clues to personality. A disproportionately bigger head is indicative of organic brain disorder. In some examinees, it may suggest an overvaluation of intellectual work. Drawings of nude human figures with emphasis upon genital organs and breasts are likely to indicate psychosexual conflicts; emphasis upon eyes and ears indicate paranoid tendencies, emphasis upon lips in male drawings tends to indicate homosexual tendencies, and emphasis upon hands tends to indicate sexual preoccupations, hostility, aggression and also the guilt-laden tendencies. Sometimes, the hands are omitted from the figure and this omission is taken to be indicative of self-punishment as a consequence of an unconscious feeling of guilt. Such omissions are mostly found in the drawings of sex offenders and juvenile delinquents. The clothing and accessories have also some meaning in drawings. The emphasis upon pockets on the shirt of male figures drawn by a male examinee is taken to be indicative of homosexuality. Holding of a knife by the figure drawn indicates aggression and hostility. Emphasis upon the belt in the drawing of a person (male or female) tends to indicate strong tendencies to control sexual impulses. Incapability to draw a figure of the opposite sex is taken to be indicative of latent homosexuality on the part of the examinee.

Toy Tests

In toy tests, toys like dolls (representing adults and children of both sexes), puppets and miniatures are given to the examinees, mostly children, who are allowed to play with the given objects in the way they desire. The examiner carefully notes down the items chosen by a child for playing, the manner in which he handles them, his emotional verbalization and other overt behaviour (Bell 1948). A child's play with these objects is expected to display his sibling rivalries, conflicts, fears, aggression, etc. One popular example of a toy test is the World test developed by Lowenfeld in 1939 and later revised and restandardized by Buhler, Lumry and Carroll in 1951, and Bolgar and Fischer in 1947. The test consists of 150 to 300 miniature pieces of objects like trees, cars, fences, houses, bridges, animals, people, etc. The examinee is requested to construct whatever he likes from these materials normally placed on a large table. The examinee notes down the objects chosen, the sequence in which the objects are made, expressive reactions like verbalizations, emotional expression and other overt behaviours. All these items of information are considered important for providing interpretative clues.

Artistic Productions

Artistic productions like drawings, finger-paintings, brush paintings, clay modelling and sculpture are some of the common projective devices through which unconscious dynamics of the examinee may be judged. It is a common observation that mentally ill persons display distortions, lack of symmetry, disproportions, stereotype and similar features in their paintings or drawings. Depicting ambitious maps, plans, projects and complex designs have been found to reveal paranoid tendencies. Emphasis upon sexual themes in painting or drawing reveals disturbances in psychosexual development. The colours chosen for painting also provide important clues to the personality.

Graphology

Graphology (or handwriting) is also considered as an important form of the expressive projective technique, which provides a clue to the personality. One of the first attempts to study personality on the basis of handwriting was made in 1622 by Camillo Baldo, an Italian physician. However, the scientific analysis of handwriting as a means of studying personality in modern times started in 1875 when J H Michon, a French scientist, published his book *System of Graphology.* This attempt to study graphology became recognized in 1896 when L Klages established the Graphological Society in Germany. Klages made a systematic analysis of handwriting of hysteria patients. Robert Sandek studied the handwriting of different persons very extensively and his attempt reached a point of culmination with the publication of his book entitled *The Psychology of Handwriting* in 1925. Since the publication of this book, a vast interest has grown among scientists for studying handwriting as a means of evaluating personality.

The study of handwriting characteristics of children as well as adults has revealed several interesting points. Children having strong self-confidence and emotional stability tend to write longer letters with strong pressure whereas bewildered and preoccupied children usually write short letters with low pressure and hesitating expression at the beginning of letters. In severely disturbed children several alternatives of the forms of the letters such as the exaggeration of the size of letters, splitting of the letters and words, etc. were found to be most obvious. Likewise, studies of adults' handwriting have revealed that persons having brain disorders show erratic and irregular flow in size of the letters because of the impairment of motor control essential for smooth writing. The writing of such persons is careless, showing frequent blotting and omission of letters. A very slow writing with a downward tendency in words or lines as well as shrinking of the initial letters of the lines displays the depressive mood of the examinee.

EVALUATION OF PROJECTIVE TECHNIQUES

Readers, by now, should have a clear picture regarding the different projective techniques. Some of these techniques such as the Rorschach test and the Thematic Apperception Test are very popular, whereas others are little known—partly because they have emerged recently and partly because they do not permit objective measurement. To evaluate separately each of the projective techniques discussed so far is very difficult, in the sense that it would require a separate chapter in itself. As such, an overall evaluation of projective techniques, in general, is presented below.

Fakability

In projective techniques, answers given by the examinees do not hurt or pose any threat to their self-image. As such, they never tend to conceal or distort their true answers. In self-report inventories, situations are more structured and direct. Before making his answers known to others, the examinee thinks a lot and ultimately moulds the true responses most of the time in the light of what is considered socially desirable or undersirable. Thus, one of the important advantages of projective techniques over the self-report measures of personality is that they are not susceptible to the fakability of responses. Despite this, there is evidence to support the fact that examinees can fake their reponses even on standard projective techniques like the RT and the TAT when they are specially instructed to modify or change their responses (Masling 1960).

Objectivity

It is said that the projective techniques are less objective than the self-report inventories. In reality, neither the projective tests nor the self-report inventories are objective. Both possess an element of subjectivity but in different ways. Projective tests are said to be subjective in scoring because scoring a response is largely dependent upon intuition and the experiences of the examiner. This automatically means that the results of any two or more examiners cannot be totally identical. Above all, the interpretation of scores in projective techniques is even more coloured with the element of the subjectivity so that quite often the interpretation reveals more about the personality of the examiner than the examinee. As a matter of fact, the interpretation for the examiner becomes as projective as the test stimuli for the examinees (Masling 1960). The self-report inventories, on the other hand, no doubt possess the trait of objectivity in scoring, but the validity of the scores obtained on these tests is largely dependent upon the subjective processess of the examinees. The examinees may not give the true answers which may invalidate the results, irrespective of the fact that the scoring was objective.

Standardization

Most projective techniques are unstandardized, that is, they lack uniformity in scoring and interpretation. The examiners have a free play in scoring and interpreting the responses obtained on standard projective techniques like the RT, the TAT and the CAT. Not only that, it has also been found that the sex of an examiner influences the types and number of responses that the examinee gives on the RT (Masling 1960). A male examiner tries to get a particular and larger number of responses from a female examinee. Self-report inventories, on the other hand, are mostly standardized and therefore, chances of free play by the examinees are the least.

Reliability

The reliability of most projective techniques is usually very low. There are probably two reasons for this. First, there is no uniform standard way of scoring and interpreting the responses obtained on projective techniques. As a consequence, scorer reliability is very poor. Second, these tests do not yield consistent scores upon retesting of the same subjects. It is also not possible to calculate the alpha coefficient (internal consistency of coefficient) of most projective techniques because the items are not usually comparable. The only logical estimate of reliability of such techniques is through the parallel-form method. Moreover, most projective techniques (except the Holtzman Inkblot test) do not have alternate forms. On the whole, the computation of reliability of projective techniques is very difficult and whatever reliability coefficient has been obtained is, in general low.

Validity

Like reliability, the validity of most projective techniques is unsatisfactory. Most of the published validation studies on the projective techniques report criterion-related validity and most of such validity coefficients are inconclusive and debatable. In reality most of the traits measured by the common and standard projective techniques such as the RT, the TAT, the CAT and the Rosenzweig P-F study are such that they require construct validity which has not been obtained.

Situational Variables

Situational variables like physical appearance of the examiner, emphasis upon certain types of responses by the examiner and changed instructions are likely to influence the responses on the projective techniques. Some examiners may have a very formidable physical appearance, which is likely to affect the examinees' capacity to imagine, think and resort to defensiveness. The overall impact of these factors is that they affect the response productivity on projective techniques. Examiners who do not possess a frightening appearance and who tend to encourage the examinee are likely to influence the responses of the examinees favourably. Not only this, it has been also shown that changed instructions influence the score on projective techniques to a great extent. For example, in one study it was demonstrated that when the Holtzamn Inkblot test was presented as an intelligence test (thus with a changed instruction), the responses of the examinees were found to be more changed than when they were obtained under the usual instruction of the test (Herron 1964).

General Applicability

Of the three general types of projective tests, the pictorial and expressive techniques have wider applicability than verbal techniques because they can be used with illiterates, small children as well as with examinees having speech defects. In general, the self-report inventories which are suited only to literates, have poor applicability, particularly when compared to the expressive and pictorial techniques.

Eysenck (1959) has summarized his viewpoint on the critical evaluation of projective techniques as mentioned below:

1. No consistent, meaningful and testable theories underlie the projective techniques.
2. No marked correlation exists between the indicators of a projective test and the intellectual abilities as meausred or rated independently.
3. There is no evidence which shows a relationship between global interpretation of projective techniques by experts and psychiatric diagnosis.
4. There are no empirical evidences for most of the postulated relationships between projective-test indicators and personality traits.
5. There is no evidence that the conflicts, motives, needs and fantasies diagnosed by the projective techniques will yield congruent results when interpreted by psychiatrists separately.
6. Ample evidences are available to show that the majority of studies of projective techniques have flaws in their methodology, are ill-designed and are full of statistical errors in analysis of data.
7. Projective techniques, in general, have poor ability to predict failure or success in different fields of life.

☑ Review Questions

1. What is meant by projective technique? Discuss the major classification schemes of projective techniques.
2. Discuss projective technique as a tool of psychological research. Assess its effectiveness.
3. Assuming some data, explain the importance of different types of scores in Rorschach protocol.
4. Evaluate the Rorschach test and the Thematic Apperception Test as tools of psychological researches.
5. Indicate the merits and demerits of the projective method of obtaining data.
6. Write short notes on:
    (a) Verbal techniques
    (b) Figure-drawing test
    (c) TAT
    (d) Projective method of collecting research data

◆

12

TECHNIQUES OF OBSERVATION AND DATA COLLECTION

CHAPTER PREVIEW

- Methods of Data Collection
- Questionnaire and Schedule
    - Wording of Questions in a Questionnaire
    - Characteristics of a Good Questionnaire
    - Order of Questions
    - Functions of a Questionnaire
    - Types of Questionnaires
    - Advantages and Disadvantages of Fixed-response Questionnaires
    - Advantages and Disadvantages of Open-end Questionnaires
    - Advantages and Disadvantages of Mailed Questionnaires
    - Advantages and Disadvantages of Face-to-Face Administered Questionnaires
- Interview
    - Types of Interviews
    - Major Functions of an Interview
    - Factors Affecting the Uses of Interviews
    - Advantages and Disadvantages of Interviews
    - Important Sources of Errors in Interviews
    - Selection and Training of Interviewers
- Content Analysis
    - Purpose of Content Analysis
    - Methods of Content Analysis
    - Evaluation and Limitation of Content Analysis
- Observation As a Tool of Data Collection
    - Purpose of Observation
    - Important Types of Observation
- Difference between Participant Observation and Nonparticipant Observation
- Rating Scale
- Types of Rating Scales
    - Numerical Rating Scale
    - Graphic Rating Scale
    - Forced-choice Rating Scale
- Other Special Types of Rating Scales
    - Q-sort Technique
    - Semantic Differential Scale
    - Behaviourally Anchored Rating Scale
    - Nominating Technique—Sociometry

- Problems in Obtaining Effective Ratings
    - Factors Affecting Rater's Willingness
    - Factors Affecting Rater's Ability
- Methods of Improving Effectiveness of Rating Scales
    - Refinement in Stimulus Variables of Rating Scales
    - Refinement in Response Variables of Rating Scales
    - Improvement in Rating Procedures
- Errors in Ratings
    - Halo Effect
    - Error of Severity
    - Error of Leniency
    - Error of Central Tendency
    - Contrast Error
    - Proximity Error
    - Logical Error
- Evaluation of Rating Scales
- Meaning and Features of Secondary Data

This chapter discusses those research tools or data-gathering devices which are very common in behavioural researches. The obtained data are of two types: primary data and secondary data. First we shall deal with different aspects of primary data and then we shall describe secondary data and its related aspects. By primary data, we mean the data that have been collected originally for the first time. There are different types of research tools and each of them utilizes distinct ways of describing and quantifying the obtained primary data. Some of the common research tools are: *Questionnaire, opinionnaire or attitude scale, interview and interview schedule, observation method, content analysis, projective test, rating scale, sociometric test, Q-sort* and *semantic differential*. Some of these data-gathering devices will be discussed in separate sections.

METHODS OF DATA COLLECTION

Based upon the broad approaches to information gathering, data are categorized as:

(i) Primary data

(ii) Secondary data

As discussed above, primary data are those data which have been collected from primary sources on first-hand basis by the researcher. For example, determining job satisfaction of the employees of an organisation, assessing the attitude of the students towards teachers and ascertaining the quality of services provided by the workers are the examples of primary data. Secondary data are collected from secondary sources such as using census data for obtaining information regarding age-sex structure of the population. Some other examples of secondary data are use of hospital records for finding out occurrence of dengue fever among different age groups, and collecting data from books, journals, newspaper, periodicals for obtaining historical and some other types of information. Thus, secondary data are collected in a second-hand way from secondary sources. Figure 12.1 provides the details of the different methods of data collection.

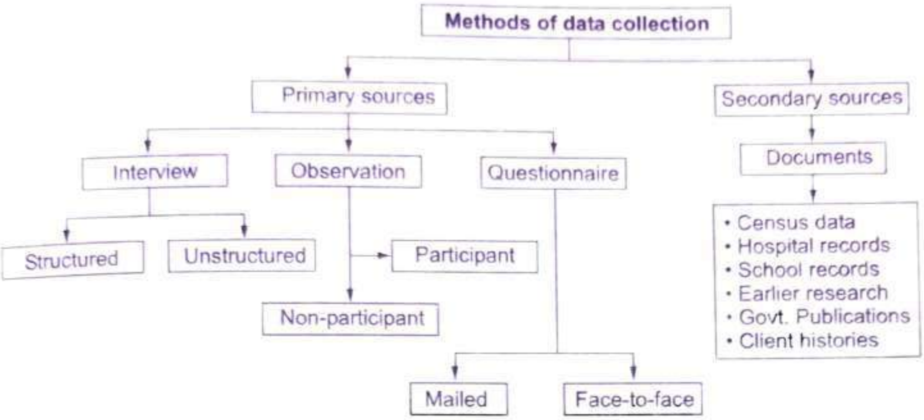

Figure 12.1: Different methods of data collection

QUESTIONNAIRE AND SCHEDULE

A questionnaire is used where factual information from the respondents is desired. It consists of a form containing a series of questions where the respondents themselves fill in the answers. A questionnaire must be distinguished from a *schedule*, an *opinionnaire* and an *interview guide*. A schedule consists of a form containing a series of questions, which are asked and filled in by the investigator in a face-to-face situation. An opinionnaire is an information form which attempts to measure the attitude or belief of an individual. Hence, an opinionnaire is also called an *attitude scale*. When factual information is desired, a questionnaire is used, but when opinions rather than facts are desired, an opinionnaire or attitude scale is used. An interview guide consists of a list of basic points or topics to be covered by the interviewer during the interview. A questionnaire is usally administered personally in groups of individuals. It has some advantages. When several persons are available at the same time and place, a questionnaire proves to be a very economical tool of data collection. Not only this, a questionnaire also enables researchers to get first-hand information regarding the vagueness of items, if any, as well as gives them an opportunity to establish a warm relationship with the persons being tested.

Wording of Questions in a Questionnaire

A vast literature on the wording of questions in a questionnaire is available. A careful review of the literature reveals that the following factors are of immense importance for any behavioural research utilizing the questionnaire as a tool of data collection.

1. *Simplicity in language*: The aim of the investigator in wording a questionnaire is to communicate effectively with the respondents in his own language. In choosing the language for the questions, he must keep in view the population for which it is meant. If the sample is taken from a general population, technical terms and jargon should be avoided. If, on the other hand, the sample is taken from a professional population, the terms appropriate to the population may be used. While one is wording the language of the questions meant for the general population, the words chosen should be informal, simple and convey the exact and same meaning to all the respondents. For example, it is more appropriate to ask: "Do you think that ......?" than "Are you of the view that ......?" or "Do you intend to think that ......?"

Investigators should, in general, avoid complex and long questions because such questions require greater efforts on the part of the respondents in answering them and many respondents may not be ready for them. Such questions are also taxing in terms of time. Although complex and long questions should be avoided, this does not mean that the short questions are necessarily the best. Marquis (1969) conducted one experiment in which the effectiveness of a questionnaire (health questionnaire) was examined as a function of the length of the questions. The same questions were prepared in short forms and in long forms. The findings indicated that the answers to the longer questions were in close agreement with reports of the respondents' physicians, thus suggesting that longer questions had a higher validity than the shorter ones.

2. *Ambiguity*: The investigator must take pains to avoid ambiguous questions because such questions do not convey the same meaning to all the respondents, and therefore, different respondents may give different answers to the same question. For example, suppose the question is: "Do you feel shy when you are in a group?" A person may feel shy when he is in a heterosexual group but may not feel shy in a homosexual group. He may, therefore, answer "Yes" or "No" depending upon his attitude. When he answers "No", what is really meant, is difficult to tell. Likewise, double-barrelled questions may introduce ambiguity and hence, should be avoided. Suppose the question is: "Do you enjoy travelling in a car and bus?" A respondent who likes travelling in a car and dislikes travelling in a bus would be in a dilemma and, therefore, is likely to be guided by his discretion. In reality, for such double-barrelled questions, two independent questions should be given so that the respondent is never placed in a state of dilemma and can answer according to his likes and dislikes. Sometimes not any particular wording can be accepted by all. In such a situation, the *split-ballot technique* is useful where different wordings of the same questions are used for different equivalent groups of respondents and subsequently, their responses are compared for knowing the effect of the wordings.

3. *Vague words*: Vague words should also be avoided because they encourage vague answers. Words like 'often', 'generally', 'fairly', 'on the whole', etc., should be avoided unless the investigator is interested in vague answers. Vagueness is also introduced by the use of 'why questions'. Lazarsfeld (1935) has discussed in detail how such questions introduce vagueness. For example, suppose the question is: "Why do you not want to educate your wards in a co-educational institution?" The answer may depend upon a host of factors, which may apply to some respondents and may not to some others. Therefore, nothing definite can be concluded about the respondents.

4. *Embarrassing questions*: When respondents are asked questions about private matters or regarding matters which they do not want to make public, they are embarrassed. For example questions relating to sexual behaviour, stealing, cheating in examinations, etc., may be embarrassing to respondents and hence, they will either refuse to answer or distort their true answers. Some investigators have suggested methods to deal with the embarrassing nature of questions so that their threat is reduced. One simple method for doing so is not to ask such questions directly from the respondents but to ask them to express their views in others. Likewise, sentence-completion techniques may also be adopted to reduce the threatening nature of embarrassing questions. Oppenheim (1966) used the technique of sentence completion for studying the attitude of the psychiatric nurses towards a mental hospital and found the technique quite effective in reducing the threatening nature of embarrassing questions.

5. *Double negatives*: Double negatives should be avoided. Such negatives tend to cancel each other and, therefore, create confusion for the respondents. For example: Do you not approve the idea that a college girl should not engage herself in domestic affairs?

6. *Leading questions*: A leading question should also be avoided. By definition, a leading question is one, which by virtue of its content and structure, leads to a specific answer. For example, the form of the question: "Should something not be done about...?" usually leads to a positive answer and the question form: "You don't think.... do you?" usually leads to a negative answer. Question builders should avoid such leading questions. Payne (1951) and Cantril (1944) have investigated the role of leading questions and have concluded that there are numerous words which, if inserted in the question form, may lead to a specific answer. For example, the word, "involved" in a question like: "Do you think that the party should get involved in....?"

generally leads to an answer in the negative. Sometimes the words leading to a specific answer may be so subtle that they go unnoticed. The questionnaire builders should take special care when using subtle leading words.

7. *Presuming questions*: Presuming questions are those which presume something about the respondent. In other words, the question should not be such that necessarily implies that the respondent possesses the knowledge regarding the theme of the question or that he has participated in those activities, which are being asked. Suppose a housewife who has never used 'nutri nuggets' (a kind of vegetable protein food prepared from soya) is asked: "Do you find nutri nuggets tasty?" This is one example of a presuming question, which should be asked only after a housewife has used the product.

8. *Hypothetical questions*: Hypothetical questions are of little value because the respondents' answers towards such questions do not reflect anything concrete. The form of the questions: "Would you like to....?" and "What do you do if....?" illustrates hypothetical questions and as far as possible, such wordings should be avoided.

Characteristics of a Good Questionnaire

A good questionnaire must have the following characteristics:

1. The questionnaire should be concerned with specific topics, which must be regarded as relevant by the respondents. The investigator must clearly state the significance, objectives and aims of the questionnaire either in a separate letter or in the questionnaire itself.
2. The questionnaire should, as far as possible, be short because very lengthy questionnaires often find their way into the wastebasket.
3. Directions and wordings of the questions should be simple and clear. Each question should deal with a single idea.
4. The questions should be objective and should not provide any hints or suggestions regarding a possible answer.
5. Embarrassing questions, presuming questions and hypothetical questions should be avoided.
6. The questions should be presented in a good order, proceeding from general to specific responses, or from those showing a favourable attitude to an unfavourable attitude.
7. Lastly, a questionnaire must be attractive in appearance, neatly printed or duplicated and clearly arranged.

Order of Questions

In a questionnaire, the order of the individual questions is of great significance because the order tends to influence the validity of the obtained answers as well as the refusal rates (Cantril 1944, Whitfield 1950). In the beginning, the respondent is unsure of himself as well as curious. So, the opening questions should be simple, general, and such as to put the respondent at ease. This has a natural effect of establishing rapport between the investigator and the respondent. No sensitive question or embarrassing question should be put in the beginning because it is likely to lead to refusal to answer. Then the questions should move from the general to specific aspect in a logical manner. Ordinarily, no break should be given in moving from general to specific questions. But where the break becomes essential, the investigator should explain the significance of the break and the new set of questions should follow in one or two sentences. Preferably, the sensitive questions which can be avoided, should be placed in or near the end so that if refusal is met, relatively only a few questions are left unanswered.

Kahn & Cannell (1975) have recommended that it is most wise to start with some broad questions relating to the topics and then, gradually narrowing down to the specific questions relating to the topics. They have named such a sequence of questions as a funnel sequence. Currently, the funnel sequence of questions has become the standard norm for questionnaires in behavioural researches.

Functions of a Questionnaire

A questionnaire performs generally two functions as given below.

1. *Description*: One of the basic functions of a questionnaire is to describe the individual or group characteristics. In other words, the questionnaire provides description about age, sex, marital status, occupation, income, political affiliation, religious affiliation, membership to some civic group or corporation, etc. These pieces of information, in turn, serve many purposes of the investigator or researcher. For example, if a researcher is able to know the age distribution of a group of employees in a university, he may be able to draw better inference and explanation for certain group behaviours such as clique formation, inter-group rivarly, etc.

2. *Measurement*: Another important function of a questionnaire is the measurement of individual and/or group variables like attitude, opinion, personality traits, etc. The questionnaire may consist of several items which aim at assessing such attitude, opinion, traits and habits of the persons.

Types of Questionnaires

Questionnaires, as used in behavioural researches, can be classified on the basis of two dimensions—(a) type of response required, and (b) type of questionnaire administration.

Based upon the type of response required, the questionnaire may be of the following two types—

1. Fixed-response questionnaire
2. Open-end questionnaire

A discussion of these two types of questionnaires is given below.

1. *Fixed-response questionnaire*: As its name implies, a fixed-response questionnaire is a questionnaire which consists of statements of questions with a fixed number of options or choices. The respondent is asked to check the option or response that best fits or suits him. Such a questionnaire is also known as *closed-form questionnaire or pre-coded type of questionnaire*. A few statements illustrating fixed-response questions are given hereunder.

(i) Do you feel shy in talking to members of the opposite sex? Yes/No

(ii) Do you like to entertain members of the opposite sex in a club? Yes/No

(iii) Do you like to have a member of the opposite sex as one of your shopping partners? Yes/No

One of the basic assumptions to be made behind the use of a fixed-response questionnaire is that the target sample has an adequate knowledge of the subject matter of the questionnaire. Another assumption is that the researcher has enough knowledge about the sample under investigation so that he can easily anticipate what kinds of responses are likely to be given.

2. *Open-end questionnaire*: An open-end questionnaire is a questionnaire which consists of questions that require short or lengthy answers by the respondents. Usually, here the answers are longer than those given in the fixed-response questionnaire. The following examples illustrate open-end questions:

(i) What are the causes of student unrest?

(ii) What methods do you recommend for improving discipline on the university campus?

There are questionnaires that are made up of items having both fixed and open-ended questions. Champion (1969) constructed a questionnaire in which both types of questions were used. This questionnaire had been administered to a group of seniors at a small college.

Based upon the method of administering questionnaires, the following are the two common types of questionnaires:

1. *Mail questionnaire*: A mail questionnaire is a questionnaire which is mailed to the designated subject with a request to answer the questions and return it through mail. Instructions for completing the questionnaire are usually enclosed and a return envelope is also provided. Generally, the researcher waits for a fortnight or so for the reply. A survey conducted in this area has revealed that about 70% of the questionnaires mailed are not returned.

2. *Face-to-face administered questionnaire*: The face-to-face administered questionnaire is one where the selected subjects are given questionnaires with instructions to complete them in the presence of the investigator or his associates. This type of questionnaire is more common than the mailed questionnaire. Face-to-face administration of a questionnaire is usually preferred where subjects for the study are readily available at one place.

Advantages and Disadvantages of Fixed-response Questionnaires

A fixed-response questionnaire has both advantages and disadvantages. The important advantages are given below.

1. Fixed-response questions or items are easily scored and coded. This facilitates the statistical calculation and helps the researcher in arriving at a conclusion soon.
2. Such questions or items require no writing from the respondents, because they are simply to check the response that applies to them. Such items or questions become more advantageous to those respondents who, due to some reason, can't adequately express themselves verbally.
3. Fixed-response questions or items usually take less time in their completion. Even lengthy questionnaires with fixed-response items can be completed rapidly in less time than those questionnaires that require written answers for the same type of information.
4. If such a questionnaire is sent to the subject by post, the researcher can expect that it will be returned by the subject soon because such questions take least effort and time in their completion.

Despite these advantages, such a questionnaire has some limitations as mentioned below.

1. One important disadvantage of the fixed-response item is that, here, the researcher remains unable to provide the respondent with all relevant response alternatives. If the respondent is forced to make a choice among several such alternatives that, in fact, don't fit him, the resulting information will be misleading for the researcher.
2. A fixed-response questionnaire sometimes encourages the respondent to adopt some kind of response set or bias. In an attempt to get over it quickly, the respondent may check only the first option and ignore others in most of the items or questions. In such a situation, again, the final information will be misleading for the researcher.

Advantages and Disadvantages of Open-end Questionnaires

An open-end questionnaire also, like the fixed-response questionnaire, has both advantages and disadvantages. The following are the major advantages of open-end questionnaires:

1. Open-end questionnaires have been found to be particularly beneficial when the investigator has little or no information about the subjects to be studied.
2. Since an open-end questionnaire provides a greater degree of flexibility, it encourages the investigator in eliciting unanticipated and insightful replies from the respondents that actually increase the researcher's understanding of what is going on and why.

Despite these advantages, there are also some disadvantages of open-end questionnaires as given below.

1. Several types of biases may operate in an open-end questionnaire. In the sample being studied, some subjects may be adept in the art of self-expression while some subjects may be poor in it. In such a situation the latter type of subjects will be unfairly combined with the former type of subjects, and therefore, an educational bias will operate which will produce misleading results. Socio-economic differences also tend to contribute to misleading results. Subjects of different socio-economic backgrounds don't necessarily see things in one and the same way, nor do they use similar vocabulary to express themselves.
2. Another disadvantage is that an open-end questionnaire is time-consuming as compared to the fixed-response questionnaire. Not only this. If open-end questionnaire is mailed to the respondents, the response rate will be lower as compared to the situation where fixed-response questionnaire is used.
3. Open-end questions are difficult to be objectively scored or coded. Different respondents may appear to provide similar responses to the same item on a questionnaire but the importance and meaning that each respondent applies to his answer may be different. Generally, an attempt is made to provide several individuals in one and the same category for the purposes of facilitating and analyzing data. The results of such an analysis are usually misleading.

Advantages and Disadvantages of Mail Questionnaires

For a discussion of advantages and disadvantages of mail questionnaires, the readers are requested to consult Chapter 15.

Advantages and Disadvantages of Face-to-Face Administered Questionnaires

In face-to-face administration, the questionnaire is administered to the people in presence of the investigator or researcher. This type of questionnaire has the following advantages:

1. A face-to-face administered questionnaire is less time-consuming. Here, the rate of questionnaire completion is high, and a better return as compared to mailed questionnaires is expected.
2. The investigator knows well who is completing the questionnaire or answering the questionnaire. This increases the validity of the collected information.
3. The researcher is present to answer any query that the respondent may raise during the completion of the questionnaire.
4. A face-to-face administered questionnaire enables the investigator or researcher to put probe questions, if any, for detailed analysis.

However, a face-to-face administered questionnaire has some limitations as indicated below.

1. The presence of the investigator or researcher makes the examinees or the respondents too conscious and that sometimes adversely affects the validity of the information thus collected.
2. If the researcher or the investigator is required to travel throughout the city or state for contacting respondents personally, the technique becomes much more costly and time-consuming as compared to the mail questionnaire.
3. In such a questionnaire administration, the group of the respondents is often selected according to accessibility and convenience. Therefore, such a group usually no longer remains representative of the population. This affects the external validity of the questionnaire.

Thus, we find that the different types of questionnaires have their relative advantages and disadvantages. Keeping in view these advantages and disadvantages, the researcher selects the appropriate ones for use.

INTERVIEW

The procedure for interview is different from that for the questionnaire, but both have the same aim, and it is to obtain data regarding the respondents with minimum bias and maximum efficiency. Interview is a face-to-face situation between the interviewer and the respondent, which intends to elicit some desired information from the latter. Thus an interview is a social process involving at least two persons, the interviewer and the respondent. For success of the interview, one must take care of the *interaction* between the interviewer and the respondent. The respondent's answer to the questions raised by the interviewer and his other behaviour serve as important clues to the interviewer and are likely to affect the behaviour of the latter. Likewise, during the course of the interview, the respondent tries to size up the interviewer and his inference about the interviewer is likely to influence his answers. Apart from these, the success of the interview is also dependent upon three important conditions, namely, accessibility, cognition and motivation, which have been discussed in Chapter 15.

Types of Interviews

There are two types of interview, namely, *formal interview* and *informal interview*. A formal interview may be defined as one in which already prepared questions are asked in a set order by the interviewer and answers are recorded in a standardized form. The formal interview is gaining much popularity today. The investigator aims at having interviews conducted in a uniform way. The formal interview is also known as a *structured or patterned interview*. Thus the formal interview is a systematic procedure for collecting information regarding the respondents and therefore, it is not surprising that the reported validities for formal interviews are usually higher than those for the informal interviews. As the interview situation is highly structured, that is, the questions, their sequence and scoring methods are all predetermined, relatively less trained interviewers can also conduct such an interview smoothly. There are two important limitations of a formal interview. First, the procedures of conducting a formal interview are expensive and time-consuming. Therefore, a formal interview is conducted only where an informal interview cannot be conducted. Second, the validities of the formal interview are usually less than those obtained by some common methods of biodata analysis and standardized psychological test (Guilford 1966).

An informal interview is one where there are no pre-determined questions nor is there any preset order of the questions and it is left to the interviewer to ask some questions in a way he likes regarding a number of key points around which the interview is to be built up. As most things depend upon the interviewer, the situation remains unstructured and, therefore, such an interview is also known as an *unstructured interview**. An informal interview is more commonly used than the formal interview and is a flexible method of collecting data. The primary advantage of the informal interview over the formal interview is that in the former, the interviewer can 'dig deeper' and thus, get a deeper understanding of the respondents' behaviour. As the interviewer is left free to ask the questions, he can mould questions in such a way that may reflect the deeper aspects of the respondents' personality. As a formal interview is relatively inflexible, the interviewer is bound to proceed with the set questions and thus, is deprived of the method of probing the respondents' behaviour. Despite these, the informal interview has three important limitations. *First*, in an informal interview there is greater scope for personal influence and bias of the interviewer as compared to the formal interview. In other words, in the informal interview there is 'more of the interviewer' than of the 'standard question' and hence, the informal interview is less reliable than the formal interview. *Second*, an informal interview requires greater skill on the part of the interviewer than the formal interview. The conduct of an informal interview requires that the interviewer be tactful, intelligent, and have a social sense as well as a deeper knowledge of the subject matter. Due to these requirements, the scope of the informal interview is limited. Third, the data obtained from an informal interview is difficult to quantify and analyze because of three inherent difficulties. One is that since interviewers are left free to ask the questions from the respondents, different questions are likely to be asked from different respondents. In such a situation, it is difficult to aggregate and summarize the results and draw any meaningful conclusion. Second, even if the interviewer asks more or less the same questions to each respondent, differences in language used by the interviewer are likely to make the responses not comparable statistically. Third, the results obtained from the informal interview are not amenable to statistical analysis. This further complicates the matter and puts the informal interview at a disadvantage.

Major Functions of an Interview

Interview as a research tool is selected basically because it serves two functions which mark it out with positive advantage from the rest of the methods of data collection. These two functions are:

1. Description, and 2. Exploration.

These two may be discussed as under:

1. *Description*: An interview has been found to be particularly useful in providing insight into the interactive quality of social life. In an interview, people spend most of the time with one another in some form of verbal interaction. The verbal interaction enables the interviewer in understanding how people view the subject under investigation. This understanding helps him know his social life which is otherwise abstract and merely a statistical phenomenon.

2. *Exploration*: Another purpose of an interview is to provide insight into the unexplored dimensions of a topic or subject. A review of the work done in this area reveals that an interview helps a lot in exploring some new variables for study as it also helps sharpen the conceptual clarity. Talking with interviewees and thereby gaining insight into their conduct from inquiries about their behavioural dimensions provides adequate stimulation for development of various hypotheses for subsequent testing and research.

Of the two functions of an interview, exploration is considered more important than description.

---

* There are also several other names which are applied to the formal and informal interview. For example 'standardized' and 'qualitative'; 'mass' and 'formative'; 'inflexible' and 'flexible'; 'extensive' and 'intensive'; 'controlled' and 'uncontrolled' are some of the common names which have been used by different research investigators for formal and informal interview respectively.

Factors Affecting the Uses of Interviews

Social scientists have recognized various factors that influence the usefulness of interviews. These can be grouped into three major factors:

(1) Characteristics of interviewers

(2) Characteristics of interviewees

(3) Nature of the problem under study

A brief description of these factors follows.

(1) *Characteristics of interviewers*: A review of the work done in this area reaveals that both the subjective and objective characteristics of the interviewers do influence the usefulness of the interview. Subjective characteristics are those which are peculiar to the individual and are extremely important where the major function of the research is exploratory. For the effectiveness of the interview, it is essential that the interviewers must have inquisitive minds so that they are capable of readily attuning themselves to the newly emerging facets of a problem. When the interviewer possesses this quality, he can readily sharpen the questions, and redirect and concentrate on other matters if needed during the time of interview. The interviewer must also have the ability for drawing together the scattered pieces of information into a uniform and integrated whole, which helps him conduct the interview in a smooth way.

Besides these subjective characteristics, objective characteristics of the interviewer are also important from the point of view of the effectiveness of the interview. The major objective characteristics of the interviewers such as sex, age, race, manners, clothing, culture, education social class, speech, etc., are important in determining the effectiveness of the interview because they interfere with, rather than influence, the verbal dialogue that takes place between the interviewers and the interviewees (Gorden, 1969).

In fact, any research problem incorporating the interview as a data-collection tool must take into consideration both these types of qualities of the interviewers.

(2) *Characteristics of the interviewees*: Since an interview is a special form of conversation, it is expected that the characteristics possessed by the interviewees can also affect the effectiveness of the interview. There are two basic characteristics of the interviewees that tend to influence the effectiveness of the interview.

(i) The first characteristic is the capacity of the interviewees to verbalize. Therefore, the very young and the mentally retarded or ill and others with extremely limited communication skills are not suitable as effective interviewees. Likewise, people with little formal training and those with relatively isolated personal conditions are not able to verbalize their views in a meaningful way and, therefore, are not effective interviewees.

(ii) The second characteristic is the willingness of the interviewees, i.e., the interviewees must be willing to verbalize their viewpoints regarding the research problems. If not, they will not express their views about the questions asked by the interviewers. Hence, they are not considered as good and effective interviewees. Lane (1962) suggested that for increasing the willingness of the interviewees to be interviewed, they should be paid money for participating in the interview and it should be conducted on the interviewee's terms.

(3) *Nature of the problem under study*: The nature of the research problem or topic also influences the effectiveness of interviewing. In general, it is believed that when the research problems are such that touch the segments of the persons' private lives and/or when the problems are such that they create special difficulties in verbalization, they tend to have an impact upon the scientific quality of interviewing (Cressey 1953; Sutherland & Cressey 1956). For example, people don't want to reveal how much money they earn from their work, as well as they don't want to verbalize the peculiarities or incidents of their sex lives. As such, interviewing about these research topics is bound to have an impact upon its scientific qualities.

Thus, the effectiveness of interviewing is determined not only by the qualities of the inteviewees and interviewers but also by the nature of the research problem.

Advantages and Disadvantages of Interviews

The interview as a research tool has both advantages and disadvantages. According to Gorden (1969), its major advantages are as follows:

(1) An interview allows greater flexibility in the process of questioning. As such, many types of probe questions can be put and analyzed.

(2) It facilitates the investigator in obtaining the desired information readily and quickly.

(3) It facilitates the investigator in being sure that interviewees have themselves interpreted and answered the questions. This increases the validity of the conclusion arrived at.

(4) In an inteview, a desired level of control can be exercised over the situation or context within which questions are asked and answers are given.

(5) The validity of the verbal information given by the interviewees can easily be checked on the basis of their nonverbal cues.

Still, the interview is not without limitations or problems. Its major *disadvantages* are mentioned below.

(1) *Interviewer's variability*: It is commonly found that at times, the interviewer views similar responses differently and records them differently from interview to interview. Thus, he himself becomes a source of variation. It has been observed that with the degree to which the interviewer becomes a source of variation from interview to interview, the effectiveness and dependability of interviewing are adversely affected.

(2) *Inter-interviewer variablity*: Inter-interviewer variablity is one of the major problems with interviewing. When several interviewers are used in a study and the nature of interview is unstructured, the interviewers vary considerably among themselves in their respective abilities to elicit the exact kind of information needed, project the proper kind of image, record the information appropriately, and so on. As the number of interviewers tend to increase, the problem of the interviewer's variablity becomes all the more compounded and affects the real purpose of the interview.

(3) *Validity and dependability of verbal responses*: In an interview, the interviewees verbally answer the questions asked by the interviewers. Social scientists have grave doubts whether a person actually behaves the way he professes to behave. They have expressed the concern that verbal responses can't be relied upon with a considerable degree of validity and dependability.

(4) *Time*: The interview takes much time in its completion because each respondent or interviewee is interviewed individually and the records of the verbal interaction of each respondent is kept individually. Sometimes, tapes are used for recording the proceedings of the interview for curtailing the time taken unnecessarily in recording by the interviewers. But even then, additional time is taken to transcribe information from tapes.

(5) *Variations inherent to the interviewing context*: Some social scientists like Cicourel (1964) have shown that the interviewing context does not ordinarily remain constant as interviewers move from one interview to another. In such a situation, the investigator remains unable to depend upon the facts obtained from interviewing. Thus, the context of the interview itself becomes a variable which must be accounted for in assessing findings from studies that utilize interview as one of the important research tools.

(6) *Recording information*: How to record information being given by the interviewee is also a problem in interviewing. No foolproof system of recording has yet been worked out to everybody's satisfaction. Some prefer to jot down the main comments or points made at crucial

junctures during interviewing, while others prefer to record the information in details, and still others prefer to rely on some type of recording device for the interview. One advantage with recording the interviews is that there is no problem with remembering the chronological sequence of the interaction taking place between the interviewers and the interviewees. The preceding two methods of recording are subjective and vague. Since there is no standard way of recording information, the dependability of the interview becomes adversely affected in the absence of a standard criterion.

Thus, there are several disadvantages associated with an interview. By making the interview structured, several of the disadvantages can be done away with, but even then little can be done to offset the limitation of time.

Important Sources of Errors in Interviews

Although the interview is commonly used as an important method of data-collecting device in behavioural sciences, some important sources of errors have been located by psychologists, sociologists and educationists. The major sources of such errors may be enumerated as follows:

(1) *Attitude of the interviewer* : One major source of error in interviews is the attitude of the interviewers themselves. Sometimes the interviewers carry a definite attitude and bias toward the favourability or unfavourability of the respondents or interviewees. Due to this type of attitude or preconceived notions, the interviewers remain unable to arrive at a correct evaluation as also they commit several errors in recording the information given by interviewees. This unnecessarily creates problems in the sources of interviewing.

(2) *Incomprehensibility of the questions asked* : Sometimes, the questions put by the interviewers are deliberately made difficult and incomprehensible. As such, the interviewees are unable to give correct answers. In such a situation, the data obtained is not dependable and the investigator is bound to arrive at a wrong conclusion.

(3) *Lack of warmth in the situation of the interview* : Gorden (1969) is of the opinion that when the interviewers don't exhibit an affectionate and warm relationship with the interviewees and, instead, exhibit an unfriendly and curt behaviour, the interviewees naturally don't co-operate with the interviewers. This becomes a major source of error in the interview, and consequently, whatever data is thus collected lacks dependability and validity.

(4) *Lack of motivation in respondents*: When the interviewees or respondents don't possess proper motivation to answer the questions asked during the interview, it constitutes a major source of error in interviewing. The data obtained no longer remains representative of the true viewpoints of the interviewees who, in this situation, mostly give their answers, as `Yes' or `No'. The investigator remains unable to arrive at any meaningful and defnite conclusion.

(5) *Duration of interview* : Sometimes the duration of the interview is unnecessarily long causing the interviewees to feel nervous and monotonous. As such, the information given by them no longer remains dependable and the interviewers tend to arrive at erroneous conclusions.

These sources of errors in interviews can be minimized by having skilled interviewers who can manage and manipulate the procedure in a nice way.

Selection and Training of Interviewers

The success of a formal as well as an informal interview depends upon the adequate selection of interviewers and upon their training. Needless to say, good interviewing ability is one of the basic criteria for any interviewer. Apart from this, for a successful interview, it is essential that the interviewers possess the following personal characteristics:

1. *Honesty*: The interviewer must be honest and scrupulous in order to get an effective interview and make a valid contribution to the research. Only when the performance of an interviewer over several interview boards is compared and assessed, we are able to know his level of honesty in interviewing practice.

2. *Accuracy*: The interviewer should be accurate in recording the answers given by the respondents. Not only this, in formal interviewing, he must also be accurate in following the sequence of the set questions. Interviewers, who commit mistakes in carrying out their administrative duties, should be avoided.

3. *Adaptability*: By adaptability is meant the capacity to adjust in varying circumstances in which the interview is to be conducted. The interviewer may have to conduct an interview sometimes with persons living in slum areas/rural areas and sometimes with people living in urban areas. In such a situation, he must possess the trait of adaptability so that he can interview these different types of respondents without showing any prejudice. Those interviewers, who have strong prejudices against some people, should not be entrusted with the task of interviewing them.

4. *Interest*: The quality of work done by interviewers is also affected by their interest. If they are not interested or least interested in the interviewing work, errors and poor performance are likely to show up frequently. The general observation is that if the interviewers are assigned the task of interviewing frequently, and if the task is a lengthy one, the interest of the interviewers is likely to deteriorate.

5. *Temperament*: The temperament of the interviewers is likely to influence the output of the interview. They should not be emotionally involved in the problems presented by the respondents nor should they exhibit characteristics such as over-aggressiveness or over-sociability because either characteristic is likely to endanger the purpose of the interview.

There is diversity of opinion regarding the training of interviewers. Some psychologists recommend a sort of formal training in which potential interviewers are given training for six to eight weeks with the help of either the lecture method or discussion method. There are some other experts who recommend on-the-job training for interviewers in which potential interviewers are asked to take their seat near the interviewers and simply watch the interviewers conducting a few interviews so that they can take over from the interviewers at a later date.

CONTENT ANALYSIS

Content analysis, sometimes known as document analysis, is a method of systematic examination of communications or of current records or documents. Instead of questioning respondents according to some scale items or observing their behaviour directly, the content-analyzer takes the communications or documents prepared by the respondents and systematically finds out the frequency or proportion of their appearances.

In content or document analysis the primary sources of data are letters, autobiographies, diaries, compositions, records, reports, printed forms, themes or other academic work, books, periodicals, bulletins or catalogues, syllabus, court decisions, pictures, films, cartoons, etc. It is the obligation of the researchers to establish the trustworthiness of these data that have been drawn. Content analysis can also be used with responses of projective tests, with all kinds of verbal materials and with materials specially produced for research problems.

Purpose of Content Analysis

The following major purposes are served by document or content analysis:

(i) To explain and describe the prevailing practices or conditions

(ii) To identify concepts, beliefs, thinking and literary styles of a writer

(iii) To locate and explain the possible causal factors related to some outcome or event

(iv) To analyze the different types of errors in students' work

(v) To locate the level of difficulty of presentation in textbooks or other similar books

(vi) To analyze the use of different symbols that tend to represent different institutions, countries or even different points of view

(vii) To find out the relative importance of some topics or problems

(viii) To make careful evaluation of bias, prejudice or propaganda in textbook presentation

Methods of Content Analysis

There are different ways or methods of content analysis. Of these various methods, Berelson's (1954) method is the most useful and has been commonly applied by investigators. His method of content analysis may be presented under the three heads given below.

Specification of the Universe

In Berelson's method, the first step is to define the universe or *U* of the content. Let us suppose that the investigator is interested in studying the effect of teachers' strictness upon classroom discipline. In this problem, classroom discipline is the dependent variable and is the universe. For testing various hypotheses, the *U* should be divided or categorized into different subparts. There can be three important categories of classroom discipline—*punctuality, attentiveness* and *completion of task* by the students. Related to each of these subcategories, there can be three hypotheses:

1. If the teacher is strict, the students will come to class in time.
2. If the teacher is strict, the students will be attentive in their class.
3. If the teacher is strict, the students will do their home task.

Thus categorization of *U* directly spells out the main variables of the hypotheses and reflects the theory and problem of the study.

Unit of Analysis

Unit of analysis refers to the measure in terms of which content analysis can be carried out. Berelson (1954) has suggested five major units of analysis, namely, *words, themes, items, characters* and *space-and-time measures. Words* are the smallest unit of analysis, although sometimes letters (shorter than words) may also be used as units of analysis. Suppose the investigator wants to know the relation between the nature of words and the comprehension level of the students. The words may be categorized as easy, medium and difficult words, and each word of the list may easily be categorized into any one of the three categories, and subsequently, their comprehension level may be determined. The *theme* is another unit of analysis, which is often expressed in terms of sentence or proposition. Usually, themes are combined into sets of themes. Most of the time, themes are complex and in such a situation it is better to avoid content analysis because the analysis is likely to yield unreliable results. Despite this, the theme is a realistic unit and is very close to the original content if it is relatively simple. Suppose the investigator wants to measure the trait of self-disclosure. For this, the theme may consist of all those statements that use 'I', 'My', 'Me', 'Mine', etc. Likewise, suppose the investigator is studying child-rearing practices in a particular culture. Further suppose that he reports, "babies are breast-fed till three years of age and after that they are made solely dependent upon solid food". The theme may be categorized as 'late weaning'. *Item* is another important unit of content analysis. It refers to a whole production made by the subject towards a given stimulus. For example, short stories written by subjects towards the TAT pictures constitute an item. Likewise, a short autobiography, a short radio programme, a television programme may also be taken as the unit of an item. For behavioural research, item is one of the most important units of analysis. According to Berelson, character and time-and-space units are not important for behavioural researches. The character unit is commonly utilized by those who do research in the field of literature. In story analysis as well (which is true in the case of projective tests), the unit of character can be used. The time-and-space unit refers to the actual physical measurement of the main content, e.g., the number of pages, number of discussions, number of photographs, number of books, etc.

Quantification

Quantification is the third important aspect in content analysis. It refers to the process of assigning numerals to the objects of the content analysis. Ordinarily, this process can be completed in any of the three ways, namely, *nominal measurement, ordinal measurement* and *rating*. In nominal measurement each object, after being assigned to a proper category, is counted. Such measurement is commonly applied to those situations where the number of objects is large. Ordinal measurement consists of ranking of objects done by subjects according to some fixed criterion. For example, subjects might be asked to rank a given set of human photographs according to the intensity of emotion expressed by them. Rating is another form of quantification, which resembles the second form. The whole reproduction or the object may be rated on several dimensions. For example, an essay written by a child may be rated on originality, fluency, organization, spelling, etc.

Evaluations and Limitation of Content Analysis

Content analysis has some merits. The important merits are as follows:

First, content analysis is applicable to a wide variety of materials such as creativity, attitude, ethnocentricism, stereotypes, curriculum changes, values, interest, religiosity, college budgets, etc.

Second, content analysis can also be used to examine the effect of experimental manipulation upon the dependent variables. If the investigator wants to study the effect of practice upon the improvement of handwriting of children, content analysis may be of no less importance than any experimental design.

Third, content analysis is also used to validate other methods of observation. Suppose one wants to validate a self-disclosure inventory. It is expected that people, in general, would not like to give personal information against which the test can be validated. But subjects can be asked some projective-type of questions and the responses can be content-analyzed. Subsequently, the test can be validated against the content-analyzed response.

Despite these merits, content analysis should be used with caution because of the complexities involved.

OBSERVATION AS A TOOL OF DATA COLLECTION

Meaning and Nature: In behavioural researches, questionnaires and interviews are very common and important types of data-collecting devices. But there are situations in which these devices can't be used meaningfully. For example, when the investigator wants to see the behaviour in natural situation and study the situation-based features of conduct, the questionnaires and interviews no more serve the purpose and some form of observation becomes indispensable.

In a broad sense, the investigators constantly observe persons' behaviour. For example, the investigators observe the behaviour of the persons in experimental situations, notice various expressions of the interviewees or respondents during the interview, watch people answering the questions or items of the questionnaire, and so on. Thus, all the investigators have some first-hand on-the-scenes contact with the persons whom they are studying. But such observations are casual by-products of the investigators, which must be distinguished from the observation used as a fundamentally data-gathering device.

Observation, as a fundamental technique of data collection, refers to watching and listening to the behaviour of other persons over time without manipulating and controlling it and record findings in ways that allow some degree of analytical interpretation and discussion. Thus, observation involves broadly selecting, recording and encoding behaviour for empirical aims of description or development of theory (Weick 1968). In fact, observation, when properly and scientifically conducted, is characterized by the following features:

1. In observation, there is a natural social context in which person's behaviour is studied. Thus, observation usually occurs in natural settings, although it can also be used in such contrived settings as laboratory experiments and simulations.
2. It captures those significant events or occurrences that affect the relations among persons being studied.
3. It identifies important regularities and recurrences in social life by comparing and contrasting the data obtained in a particular study with those obtained in the study of various natural settings.

These characteristics are such as make the scientific and fundamental observation distinct from casual and are more or less spontaneous observations made by researchers in course of conducting the investigations.

Purpose of Observation

Basically, observation as a tool of data-gathering device, has the following three basic purposes:

1. One major purpose of observation is to capture and study human behaviour as it actually happens. It helps in snapshot comprehension of the activities of the persons in real life or social life.
2. Another purpose of observation is to provide a *graphic description* of real life that can be acquired in other ways. There are so many areas of life about which we have few thorough descriptions and much is taken as granted about those areas by social scientists. For example: How does a delinquent steal a motorcycle? How does a person actually go about learning to be an engineer or doctor or politician or professor? The descriptive base for all such life events is often provided only by observation.
3. Another purpose of observation is *exploration*. When the investigator observes human behaviour in a real-life setting, he gets a good chance to explore those variables which were important but overlooked. He also develops a tendency to look beyond what is already known about the subject and to examine the probability of some alternative directions for research. Not only that, observation also aims at correcting some methodological errors which otherwise might have been overlooked.

Thus, observation serves many useful purposes of which description and exploration are more important.

Important Types of Observation

There are several ways of classifying observation. On the basis of the ability of observational data to generate useful and researchable information, Reiss (1971, 4) divides observation into the following two types:

(1) *Systematic observation*: Systematic observation is one which is done according to some explicit procedures as well as in accordance with the logic of scientific inference. A psychologist studying the aggressive behaviour of children in their play group, with some objective and explicit principles decided beforehand, is an example of systematic observation.

(2) *Unsystematic observation*: Unsystematic observation is a type of casual observation made by the investigator without specifying any explicit and objective inference. A psychologist or sociologist observing the behaviour of people on a railway platform without any explicit principles and procedures is an example of unsystematic observation.

Observation has also been classified on the basis of the role played by the investigator. On the basis of this criterion, observation may be classifed into *Participant observation* and *Nonparticipant observation*. A discussion of these two may be presented as under:

(A) Participant observation: As its name implies, in participant observation the investigator actively participates in the activities of the group to be observed. Here, the investigator may already be the member of a group or organization and decide to observe it under one or more situations. Or, he may join the group for the express purpose of observing the group under one or more situations. The procedure of participant observation is often unstructured and, usually, the identity of the observer is not known to other members of the group. This is called *disguised* participant observation. But sometimes the persons who are being observed know that the observer is present for collecting information about them. This is known as *undisguised* participant observation. Therefore, other members of the group take him as an ordinary member and interact with him in a natural way. Since the procedure here is usually unstructured, the observer has some flexibility in deciding what to observe and how to record it. Participant observation is usually used to provide descriptions that otherwise would be unavailable. For example, in one of the most important and provocative studies by Rosenthan (1973), disguised participant observers in a psychiatric hospital posed as patients and later provided a good account of the experiences.

However, participant observation has both strengths and weaknesses. An account of its strengths is given below.

(1) In participant observation, since the observation is done in a natural setting, the investigator is able to record the behaviour in a realistic manner and naturally, then, the analysis yields meaningful and convincing conclusion about human behaviour.

(2) Usually, the complete observation by the method of participant observation takes several days and sometimes several months. As a consequence, whatever information is collected is very broad and meaningful for understanding human behaviour.

Despite these strengths, participant observation has some limitations or weaknesses as given below.

(1) Since participant observation is usually unstructured, it fails to be precise about the procedures for data accumulation. According to Reiss (1971), in participant observation less attention is paid to precision and more to discovery.

(2) Participant observation is a time-consuming device and, therefore, not all observers become ready to proceed by the procedures of participant observation.

(3) Since the observer participates in the activities of the group in an active manner, he sometimes starts showing human weaknesses like love, sympathy, hatred, etc., towards the members and their behaviour. This considerably jeopardizes the validity and dependability of the observation.

(B) *Nonparticipant observation*: Nonparticipant observation is the observation in which the investigator observes the behaviour of other persons in a natural setting but does not remain a participant in the activities being observed. Nonparticipant observation is usually structured, and therefore, the observer preplans the likely nature of the natural setting, representativeness of data, problems associated with the presence of the investigator, etc. Here, the observer or the investigator is able to go into the development of exploratory strategies or some specific research questions for probing.

Nonparticipant observation also has some strengths and weaknesses. Its major points of strengths may be summarized as follows:

(1) Since nonparticipant observation is usually structured, the obtained data is more reliable and representative. The observer clearly plans the different aspects and processes of observation in a nice way.

(2) In nonparticipant observation the observer is able to concentrate upon any specified aspect of social behaviour in a better way and, therefore, gets a better opportunity to find out the solution of the related probe.

However, nonparticipant observation has some limitations as mentioned below.

(1) In nonparticipant observation, the behaviour of the persons being observed and the settings do not remain a natural one. The persons develop the consciousness that their behaviours are being observed. This consciousness slightly distorts the natural flow of their behaviour. Since settings are structured, it also affects the persons being observed. But this limitation is not considered very serious for want of evidences to the contrary. So far as the present knowledge goes, there is little evidence to show that intervention in actual social context, as practised in non-participant observation, creates any problem. Likewise, there is no evidence that the presence of a nonparticipant observer tends to have any detrimental effect upon the behaviour under study (Black & Champion 1976).

(2) Nonparticipant observation fails to capture the natural context of social settings to the extent participant observation is able to capture.

DIFFERENCE BETWEEN PARTICIPANT OBSERVATION AND NONPARTICIPANT OBSERVATION

Participant and nonparticipant observations are the two different forms of observation. These two types of observation do differ as indicated below.

(i) Although both participant observation and nonparticipant observation are done in natural settings—in the former the observer or investigator actively participates in the activities of the group of persons being observed, whereas in the latter such active participation does not occur.

(ii) Participant observation is usually unstructured, whereas nonparticipant observation is usually structured. Since participant observation is unstructured, the observer has a greater degree of flexibility in deciding what to observe and how to record it.

(iii) In participant observation, the identity of the observer is often hidden and he is treated like an ordinary member of the group; but in nonparticipant observation the observer is usually known to the persons being observed. Therefore, there is little chance of concealing his identity.

This is how participant observation differs from nonparticipant observation.

RATING SCALE

There are two general classes of behaviour observation—*observation of actual behaviour* and *observation of remembered behaviour*. In the former, actual behaviour is observed, that is, the persons or objects engaged in producing the behaviour remain physically present and interact with each other. A group of college students solving a problem and the teacher–pupil interactions are examples of the actual behaviour and its observation is known as the observation of the actual behaviour. In remembered behaviour, the persons or objects engaged in producing the behaviour do not remain physically present. However, they are symbolically represented. For example, a person may be asked to recall the scene of a classroom in which the teacher is extremely dominating. Such observation is called observation of remembered behaviour. The observation of actual behaviour is easier than the observation of remembered behaviour because in the former, things are ready-made and the observer has simply to make a decision on the basis of what is around him, whereas in the latter, the observer has to take a decision on the basis of his previous experiences and/or on the basis of his ability to perceive what an object looks like (because it is not present physically). This is why remembered behaviour is also known as perceived behaviour. Rating scale is a technique to assess both actual behaviour as well as remembered behaviour.

A rating scale is defined as a technique through which the observer or rater categorizes the objects, events or persons on a continuum, represented by a series of continuous numerals. The purpose of a rating scale is to know what kind of impressions the objects or persons have made upon the raters. It is, therefore, essential that the rater must have experience or knowledge of those objects, events or persons. The experience may be direct (actual behaviour) or may be indirect or remembered (perceived behaviour). A rating scale usually has two, three, five, seven, nine or eleven points on a line with descriptive categories at both the ends followed sometimes with a descriptive category in the middle of the continuum, too. An illustration is given as follows:

Positive End:	Middle:	Negative End:
Strongly agree	Neutral	Strongly disagree

The rating scale has two components, namely, the stimulus variable and the response options. The stimulus variable (*S*) consists of trait names or qualities to be rated and the response options consist of numerical or descriptive categories. The rater or observer who is often a trained person, rates the objects, events or persons on the given scale according to his impressions and thereby, provides a quantitative analysis of his observations. Ratings may be either retrospective or concurrent. When they are retrospective, they tend to summarize all the impressions gathered by the raters regarding the ratees over an extended period of time. But when they are concurrent, they tend to summarize the impressions that are gathered as it happens in the case of an interview.

The rating scales are very easy to construct and use. As such, a large number of rating scales have been constructed. A problem which has been frequently noticed in the use of rating scales is that the traits to be rated are not distinctively defined and may have different meanings for different rates. This defeats the very purpose of ratings. To solve this problem, three precautions should be taken in the construction of rating scales. *First*, each trait to be rated should be clearly defined and explained with specific instances. *Second*, various intervals or points on the scale should be clearly defined. Usually, five to seven intervals are used in rating a trait, attitude and other sentiments. When there is need for intervals greater than seven, there is likely to be a problem of finer refinements of distinctions and in this task, even the trained rater is not likely to succeed. Hence, the intervals on the rating scale should be kept below seven. *Third*, since overt traits like leadership, honesty, punctuality, co-operativeness, industriousness and the like are more reliably rated than covert traits such as ego-strength, job-satisfaction, emotional stability, etc., attempt should be made to ensure that rating scales should, as far as possible, be concerned exclusively with objectively observable traits.

TYPES OF RATING SCALES

There are different types of rating scales. According to Guilford (1954, 263), rating scales are divided into six categories, namely, numerical scales, graphic scales, percentage rating, standard scales, scales of cumulated points and forced-choice scales.

A detailed discussion of these types is presented below.

Numerical Rating Scale

Numerical rating scales are the easiest to construct and apply to the objects, persons, events, etc., to be rated. In a numerical scale the observer or the rater is supplied with a sequence of numbers, which is well defined, and his task is to rate the objects on the given sequence of defined numbers on the basis of his impression. Sometimes, it is found that numerical scales have only a description of the categories and no numbers are provided. After the rating by the observer, the investigator assigns numerals to certain categories e.g., 5 to 'Strongly agree', 4 to 'Agree', 3 to 'Indifferent', 2 to 'Disagree' and 1 to 'Strongly disagree'. Sometimes a still more fine discrimination is needed in such a scale. Examples of such types of numerical scales are given below.

Scales with Numerical Anchors:

Numerical anchors	Meaning
1	Extremely disagree
2	Strongly disagree
3	Moderately disagree
4	Mildly disagree
5	Indifferent
6	Mildly agree
7	Moderately agree
8	Strongly agree
9	Extremely agree

Usually, the above numerical anchors along with their meanings are printed on the first page with appropriate instructions and on the subsequent pages, a number of statements revealing various impressions, attitudes, etc., regarding the objects, persons and events are pointed. Opposite each statement is provided a blank space where the rater writes simply that number which he thinks to be the most appropriate one. A few examples of such items are given below.

1. Nationalization of the private sector would make the country richer. ☐
2. Nationalization enables the government to make sound policies regarding the development of the country. ☐

Scales without Numerical Anchors:

Strongly agree
Agree
Indifferent
Disagree
Strongly disagree

When scale points are without numerical anchors, each statement is provided with all the descriptive cues where the rater puts a tick mark only on that cue which he thinks is the most appropriate. For example,

1. More and more countries should join the UNO.
   Strongly agree Agree Indifferent Disagree Strongly disagree

2. Political leaders should obey the directions of the UNO.
   Strongly agree   Agree   Indifferent   Disagree   Strongly disagree

Some of the numerical scales are assigned a rating of '0' to the neutral point +3, +2 and +1 for the different categories towards the positive ends and −3, −2 and −1 towards the negative ends. An example is given below:

+3 Strongly agree
+2 Mildly agree
+1 Agree
0 Indifferent
−1 Disagree
−2 Mildly disagree
−3 Strongly disagree

Like the first subtype, this scale is also printed on the first page and statements are given on the following subsequent pages. Opposite each statement, there is a box in which the rater writes that numerical anchor which he thinks to be the most appropriate one.

Of these three subtypes of numerical scales the first subtype is very common and advantageous, too. When the numerals are provided along with different categories of scale points, this may represent equal intervals in the mind of the rater and their ratings may come nearer to the interval measurements which, in turn, may facilitate some common statistical analysis. In the case of the second subtype, this advantage is not obvious because numerals are provided by the investigator after rating has been done by the raters. The general demerit of the third subtype of numerical scale is that its scale points are described by negative numbers such as −1, −2, −3, −4, and so on. Raters having a mathematical background may not be puzzled but those who have no such knowledge are likely to be bewildered by the negative numbers. Moreover, such scales have a point, '0', in the middle, which may suggest to some raters that the scale has a break in the middle and thus, may endanger the very continuity of the scale. For these reasons, Guilford (1954, 264) opines that the use of negative rating numbers is not recommended. The general demerit of the numerical rating scale is that the rater has to remember the meanings of the numerical anchors (which are given only on the first page) while he is writing the numbers in the box. The most common observation is that he forgets the meaning of the number and sometimes writes a wrong number in the box. To avoid this difficulty, nowadays most numerical rating scales are being designed in such a way that on each page the numerical anchors along with their meanings are given.

Graphic Rating Scale

The graphic rating scale is the most popular and widely used rating scale. In one way, the graphic scale may be considered as an improvement over the numerical scale because it tends to overcome some of the difficulties faced with the numerical scale. On the graphic scale, the scales are presented graphically in which descriptive cues corresponding to the different scale steps are given. Items or statements here have no blank box and the rater simply puts either a tick mark or a cross mark on any of the descriptive cues to indicate his view. An example of the graphic scale is presented below.

(a) Mrs Shukla delivers lectures in the classroom:

Extremely rapidly	Tolerably rapidly	With slow speed	Sluggishly	Extremely slowly

(b) In social gossip, she:

-------	-------	-------	-------	-------
Talks a great deal	Talks easily	Talks when absolutely necessary	Prefers to listen	Abstains from talking

In the above two examples, the graphic scales have been illustrated. The first example illustrates the scale points through a continuous line whereas the second example illustrates the scale points through broken lines. In both the examples, the scale points have been demonstrated in a horizontal manner. They can, however, be demonstrated in a vertical manner as well. Graphic scales having scale points arranged horizontally have one important limitation, that is, they provide space for only the shorter descriptive cues. In vertical scale points, this limitation is automatically overcome as the investigator can put as many descriptive cues as possible (because there is no limitation of space). The Fels Behaviour Rating Scale is a good example of the vertical graphic scale though it has some nongraphic scales, too.

There are some advantages of graphic scales. *First*, they have no numerical anchors. As such, the rater experiences no confusion arising out of the need for numerical discrimination. *Second*, they are simple, easily administered and quickly completed by the raters. Regarding their disadvantages, it is said that graphic scales generally take time and labour in scoring.

Percentage Rating

Percentage rating is done whenever the investigator wants a quick rating with maximum uniformity from rater to rater. The technique requires the rater to place the ratees among different specified percentage groups or into different percentiles or quartiles such as given below:

Highest 5 per cent
Second highest 5 per cent
Highest 25 per cent excluding the top 10 per cent
Top half but not the top 25 per cent
Lower half

Percentage ratings are common among teachers who are asked to rate their students in the classroom in terms of overall performance. One of the serious limitations of percentage ratings is that the rater may be quite generous and therefore, the rating may be influenced by the individual differences in generosity among the raters.

Standard Scale

Standard scales are not very popular rating scales for psychological measurement. Standard scales were originally developed in a bid to have more uniformity of meaning of scale points and to get an unbiased rating on the basis of those scale points. A standard scale is one in which the rater is presented with some standards with pre-established scale values. These standards usually consist of objects of the same kind, e.g., they may all be the names of persons. As an example the Man-to-Man Scale and Portrait Matching, which are based upon the principles of the standard scale, are given below.

The Man-To-man Scale was developed during World War I and it used men instead of numbers and adjectives or other descriptive cues to represent the various scale points. The rater may be asked to give the name of a person who is well known to him and who is very high on the trait being rated. That person's name is then noted down to define the 'very high' point on the scale. Likewise, the rater may then be asked to give the names of four other persons who are known to him and who are high, average, low, and very low on the trait being rated. All these names are entered to define the four scale points—'high', 'average', 'low' and 'very low'. Thus, the scale with its five scale points—very high, high, average, low, and very low—is complete. The rater is then given only those key names with pre-established five-point scale values to rate other persons. Here the rater's task is very simple because he is required to compare the new person or persons (to be rated) with the five key persons (defining the five levels of scale points) on the trait in question. The value of rating corresponds to the value of that key-person whom he resembles. For example, if a person is being rated on the trait of dominance and if the rater assigns him the name 'Col. Shamsher', the value of the rater on the 5-point scale will fall in 'high' on the trait of dominance. A hypothetical example to illustrate the Man-To-Man Scale is given below.

Scale Values	Name
Very high	Lt Chima
High	Col. Shamsher
Average	Brig. Kariappa
Low	Lt Bodra
Very low	Lt Gilani

There are several advantages of the Man-To-Man Scale. *First,* such a scale avoids the confusion arising out of abstract numerical anchors assigned to the traits to be rated. *Second,* if all the raters use the same key-man in their ratings, their ratings can be comparable both in absolute terms as well as in relative terms. *Third,* since the scale values of all the key-men are pre-established and fixed, raters can't shift over day-to-day's ratings. Three principal disadvantages of the Man-To-Man Scale have also been suggested. *First,* the distance between the key-men of the scale is not equal. *Second,* actual practice has demonstrated that no two raters are alike in rating persons who are well known to them and this naturally endangers the second advantage discussed above. *Third,* deliberate overestimation and underestimation of persons by the raters are not controlled by the scale.

Portrait matching is another rating scale which is based upon the principle of standard scale. The portrait matching technique was developed by Hartshrone & May in 1929 in the course of their study on character. In this technique, a set of standards (or verbal sketches or portraits) for any given trait on which rating is to be done, is prepared. For constructing the verbal portrait regarding a given trait, a large number of verbal statements describing that trait are collected and each is written on a separate card. Subsequently, they are rank-ordered by a group of judges or experts. After that, a desired number of sketches or portraits are prepared on the basis of those statements, which have about the same average rank. The scale value of each portrait is also determined. For determining the scale value, the portraits are given to another group of judges for ranking them and the mean rank becomes the scale value. In this technique the rater is given all the verbal portraits to read (which consist of several synthesized statements regarding a particular trait) and then, he names the persons who belong to the portraits. The chances are that within a single portrait, several persons and a single person may be placed under more than one portrait. A person's final rating is the average of all the portrait values that have been assigned to him by all the raters. The merit of portrait matching is that it eliminates the element of subjectivity, which is common in the Man-To-Man Scale but the most serious demerit is that it does not provide objective and realistic standard scale points. The portrait, matching technique deviates from the usual rating method in that it requires the rater to find persons to match the description (or portrait), which is in contrast to the usual practice of asking the rater to find descriptions to match the persons. This automatically means that all names of the individuals to be rated should be placed before the raters and they should not be allowed to depend upon their ability to recall the names of individuals. This is certainly an arduous, if not impossible, task for the investigator.

Scale of Cumulated Points

Rating scales based upon cumulated or summated points are the most common. Here the person's total score is the sum of individual ratings or points assigned to all items of the scale. Such points may be weighted or unweighted. At a glance such rating scales may seem identical with the psychological tests (where a person's total score is the sum of the individual scores on the items of the test) but in reality, they differ in that psychological test items are answered by the persons themselves whereas the rating scales items, which are based upon summated ratings, are answered by another person, that is, by the rater or observer. Rating scales based upon summated points can be conveniently classified into two types, namely, the *Checklist* and the *Guess-who technique*.

The Checklist method is one where the rater is supplied with a large number of specific behavioural statements (which serve as stimuli) and he is asked to check these statements, which describe the persons in question. The person (the ratee) is then characterized by only these statements which have been checked for him. Such a rating scale is known as a *behaviour checklist*. Subsequently, these assigned statements are scored. For example one very convenient way of scoring the checked statements is to score each favourable statement as +1, each unfavourable statement as −1, and each neutral item as 0. The person's total score on the checklist is equal to the sum of the scores for the items checked for him. The purpose of the behaviour checklist is to know whether certain specified traits or behaviours are present or absent in the individual being rated. A good example of the behaviour checklist is the Vineland Social Maturity Scale, which consists of items relating to self-direction, self-help, communication, socialization, etc. The checklist is filled in by the rater who knows well the children being rated.

There are several variations of behaviour checklists. The items of the checklist may be in the form of two-point responses (Yes–No, True–False) or in the multiple-choice form. The sample illustrating these two varieties is given below.

Sample item of a two-point response:

The picture was interesting Yes No

or

The picture was boring Yes No

Sample item of a multiple-choice response:

The girl was

——— smart
——— intelligent
——— beautiful
——— ugly
——— submissive

The second sample item contains only the adjectives as the response option and the stimulus (the girl) is to be rated on those adjectives. Such a behaviour checklist is, therefore, also known as the *adjective checklist*. On either of the above two types of behaviour checklists, each positive rating is given a score of +1 and each negative rating is given a score of −1. In order to get a total score, all the positive scores are summed together and from them the sum of all the negative scores taken together, is subtracted. Thus the resultant total score represents the positiveness of the rater's attitude.

In fair evaluation of the behaviour checklist it can be said that despite its simple technique and attractive procedure, it has not been a very widely adopted method. One of the serious drawbacks of the checklist is that the rating is influenced by various kinds of response biases of the rater who is simply asked to check the items in an all-or-none fashion (Guilford 1954).

The Guess-who technique, also known as Casting Characters, is another type of rating scale where the total score is based upon the cumulated points. The technique was developed by Hartshrone & May in 1929 to be used primarily with children. The technique consists of verbal descriptions (or verbal portraits) of the various roles played by children in a group. The verbal descriptions are usually in the form of one or two sentences. The child raters are asked to name the other children (or their peers) who fit or match certain verbal descriptions, mentioning the same child as many times as they think appropriate. Each favourable description is given a point and all such points are summed to get a total score. A few sample items of this technique are given below.

Here is one who is always worried.
Here is one who is always happy.
Here is one who is always discouraging others.
Here is one who is always helping others.
Here is one who never likes to do a task.

The dangers involved in using the Guess-who technique are similar to those of the portrait-matching method. The technique obviously deviates from the normal practice of asking the rater to choose the verbal description which matches the individual (because here the rater is to find the individual to match the verbal description). This technique, thus, obviously requires that the rater be supplied with names of all the individuals (or children) to be rated, and the rater should not be left to depend upon his ability to memorize the names of the individuals. This is an almost impossible task for the investigator.

Forced-choice Rating Scale

A forced-choice scale represents a major departure from the pattern of rating scales discussed above. All the above techniques consider one attribute at a time and place the ratees in any one of a set of categories. In the forced-choice rating scale, the rater is given a set of attributes in terms of verbal statements for a single item and he decides which one or ones represent the individual being rated most appropriately and accurately. The items of the forced-choice scale may have several alternatives—two, three, four or five. Of these, the two alternatives and the four alternatives are the most common. In the two-alternative form, both the statements regarding the attribute are either favourable or unfavourable. However, only one of the statements regarding the attribute in either case is valid to identify desirable or undesirable attributes, though both of them may appear equally favourable or unfavourable to the rater. In the four-alternative forms, two varieties are common—in one variety all the four statements are desirable or undesirable, and in the second variety two statements are favourable and two statements are unfavourable. When all the four statements are favourable or unfavourable, it does not mean that all of them have equal discriminative value, or in other words, that they are equally valid in identifying desirable or undesirable attributes though they may appear to be so to the rater. This automatically reduces the probability of operating a favourable response bias by the rater. The rater may be asked to select any two statements which are most descriptive and representative of the persons being rated. An example of the four-alternative forced-choice rating scale is given below.

Miss George
(a) lectures with confidence
(b) keeps the students interested and motivated
(c) cares a great deal for the slow learner
(d) entertains suggestions from students to improve her lecture

In the above four statements (all favourable), the most discriminating statement is (c) and the least discriminating statement is (a). The most discriminating statement may be assigned a score of 2 and the least discriminating statement a score of 0. A ratee's total score would be the sum of such scores assigned to the most discriminating statements in each set. The students must note carefully that the assignment of score value to the most and least discriminating statement is not random and/or dependent upon the will of the investigator. As a matter of fact, it is determined on the basis of some preliminary investigations. A comparative study of the forced-choice technique having all favourable statements and having two favourable and two unfavourable statements has revealed that the former is good for two reasons. First, the rater fully co-operates with such a form and second, the probability of faking or distortions is low whereas the same probability is high in the case of the latter type of form, that is, having two favourable and two unfavourable statements (Freeman 1962).

There are several advantages and disadvantages of the forced-choice rating scale. Its main advantages are given below.

1. The forced-choice rating scale minimizes the generosity (or kindliness) error. Since the items of the scale are equally favourable (or unfavourable), a rater who may be very kind and generous to the ratees, is not given any opportunity to choose one rather than other and thus, his tendency to rate generously is automatically controlled.

2. The tendency on the part of the rater to be influenced by a single favourable or unfavourable trait, which subsequently colours his whole judgement, is also automatically controlled. If the rater wilfully wants to bias the score in any one direction (either downward or upward) he is least allowed to do so because in the forced-choice scale, the rater is usually unable to identify the favourable choice or the unfavourable choice (because the options appear to be equally favourable or unfavourable and therefore, he is unable to rate a person up or down).

3. When all the statements of the forced-choice scale are equally attractive, the rater's bias is automatically reduced and this has a very marked effect upon the distribution of ratings, that is, the distribution of the ratings becomes symmetrical and normal.

The main disadvantages of the forced-choice rating scale are given below.

1. Raters are critical and hesitant to do ratings on the forced-choice scale because this gives them limited freedom in ratings and imposes too many restrictions upon them.

2. Sometimes, the raters believe that none of the statements actually describes the person appropriately even though they have to give their judgement. Obviously, such judgement, then, would not be representative of the rater's true decision.

Although the forced-choice scale has acquired much prominence in past years, its importance today has been considerably reduced partly due to the rater's hesitant and critical attitudes and partly due to the difficulties involved in constructing the scale.

OTHER SPECIAL TYPES OF RATING SCALES

Under this heading, some special types of rating scales, which have recently acquired prominence in the field of measurement, will be discussed. The Q-sort, the semantic differential scale, the behaviourally anchor rating scale, sociometry, etc., are a few examples. A brief discussion of these special types of rating scales is presented below.

Q-sort Technique

The Q-methodology was originally devised by William Stephenson in 1953 for the study of verbalized attitudes, preferences, self-description, etc. The procedure of the Q-methodology is known as the Q-sort, which involves a comparative rating method rather than an absolute rating method. In fact, it is a method of ranking attitude or judgements and is particularly effective when the number of items to be ranked is large. In the Q-sort technique some objects such as verbal statements, pictures, words, phrases, etc., are given to an individual or a group of individuals to sort out into a set of piles (usually nine or eleven piles) according to a fixed criterion. A limited number of piles is thus an example of forced distribution of sorting. Each object (whether it is a verbal statement or a picture or a word, etc.) is printed on a separate card. For statistical facility and reliability, the number of cards in any Q-sort studies should not be less than 60 and more than 140 (Kerlinger 1973). The sorter(s) is (are) instructed to sort a fixed number of cards into each pile so that the resulting whole distribution may approach normal or quasi-normal distribution. The purpose of approaching a normal distribution of cards, among the different piles is to facilitate statistical calculation. As Nunnally (1970, 449) has said, "The use of an approximately normal distribution rather than some other fixed distribution (e.g., a rectangular distribution) is justified in the general sense because (1) so many things in nature are distributed approximately that way, and (2) it fits in with the statistical methods applied to the data." One of the basic principles of Q-sort is that as far as practicable, the items or contents should be homogeneous. This is because in Q-sort, the investigator seeks precise comparative responses elicited by a large number of stimuli and if all these stimuli are not homogeneous, the comparative responses would make no sense. Two examples of the fixed distribution (a kind of normal distribution) are cited below to illustrate the points.

Q-sort distribution of 100 items (pictures, drawings, words, etc.):

2	4	8	12	14	20	14	12	8	4	2
10 Most preferred	9	8	7	6	5	4	3	2	1	0 Least preferred

Q-sort distribution of 90 items (verbal statements)

3	4	7	10	13	16	13	10	7	4	3
10 Most preferred	9	8	7	6	5	4	3	2	1	0 Least preferred

From both the preceding illustrations, it is obvious that Q-sort requires a rank-ordered continuum ranging from the 'Most' to 'Least' dimension with varying degrees of subdimensions in between these two extremes. The numbers above the line indicate the number of cards to be sorted in each pile. The numbers below the line indicate the total number of piles into which the objects are sorted and the values to be assigned to each card placed into different piles. For example, `10' indicates the 'most preferred' picture, and 2 pictures (the number above the line) are to be sorted in this pile and both of these pictures would be given the value of 10 and 10. Similarly, 9 indicates the next pile and 4 pictures are to be sorted into this category, each receiving the value of 9. The centre pile (5th category) is a neutral pile and the sorter is instructed to sort all those objects, which are either vague or difficult to be sorted into any one pile, into this neutral pile.

When sorting is over, statistical analysis is done. Two statistical analyses are common: the analysis of variance and the coefficient of correlation are computed between the sorting of the different sorters. The Q-data can be analyzed for one individual only or the data for the group of individuals can be analyzed. The computation of the coefficient of correlation has been illustrated from hypothetical data presented in Table 12.1 (10 pictures sorted by 5 persons into 11 piles).

Since there are five sorters, there would be

$$\frac{N(N-1)}{2} = \frac{5 \times 4}{2} = 10 \text{ pairs of sorters}$$

and accordingly, there would be 10 coefficients of correlation. Table 12.2 presents the correlation matrix (value of Pearson r) prepared on the basis of Table 12.1.

Table 12.1 Q-sort values of five sorters

Items*	Sorters				
	A	B	C	D	E
1	1	1	8	7	1
2	2	2	5	5	2
3	5	4	9	8	4
4	0	0	10	10	0
5	4	3	11	10	4
6	6	5	0	10	5
7	3	3	11	11	3
8	3	3	10	10	3
9	4	3	10	10	3
10	3	4	9	10	4

* In practice, the number of items to be sorted is far greater than 10.

Table 12.2 Correlation matrix from the Q-sort values of 5 sorters presented in Table 12.2

	A	B	C	D	E
A	+	0.932	− 0.343	0.264	0.939
B	0.932	+	− 0.363	0.292	0.978
C	− 0.343	−0.363	+	0.329	− 0.293
D	0.264	0.292	0.329	+	0.318
E	0.939	0.978	− 0.293	0.318	+

From Table 12.2, it is clear that the correlation matrix is diagonally symmetrical and hence can be read either on the basis of the upper diagonal or on the basis of the lower diagonal. Of 10 coefficients of correlation, three coefficients are very high and the remaining are below the moderate level. The high coefficients are 0.932 (between A and B); 0.939 (between A and E); and 0.978 (between B and E). This means that three sorters, namely, A, B and E, sort the pictures in the same way. Now, to know their exact way of sorting, the investigator is again required to go back to the Q-sort data and examine whether they preferred lower piles or the upper piles.

Kerlinger (1973) has divided Q-sorts into the following two principal categories.

Unstructured Q-sorts

An unstructured Q-sort is based upon the random sampling of content or materials (Nunnally 1970). The random sampling of content is different from the usual technique. Here the investigator selects a wide variety of representative contents or materials (for sorting) from different sources such as photographs of the different statues collected from different books, magazines, etc., different statements about psychotherapeutics' progress in clinics, statements about mental illness, etc. No theory, variable or specific factor is assumed to underlie those contents. An attempt, however, is made to keep all the contents or items for any one Q-sort homogeneous as far as practicable. For example, all the different photographs and/or drawings may be placed together (excluding other contents like statements, words, phrases) for sorting. Thus, in an unstructured Q-sort, the items or contents are selected to represent one domain and no specific theory or variable or factor is said to underlie those items. Moreover, such items are not partitioned further into Q-sort analysis. The principal analysis focuses on the correlations among persons on factor or cluster analysis.

Structured Q-sorts

A structured Q-sort is one in which all items, like in an unstructured Q-sort, belong to a single domain. However, these items are said to constitute a variable or hypothesis and during analysis, they can be divided into one or more than one method. These two characteristics make the structured Q-sort different from the unstructured Q-sort. In a structured Q-sort, the investigator stipulates the contents or items in terms of an experimental design. When the design is such that it permits the partition of items on the basis of a one-variable classification, it is referred to as a *one-way structured Q-sort*, and when the design is such that it permits the partition of items on the basis of a two-variable classification, it is referred to as a *two-way structured Q-sort*. An illustration of a one-way structured Q-sort is as follows: Suppose the investigator decides to make a Q-sort study of attitudes towards nationalization. He might select some statements (say 60 statements), revealing attitude towards nationalization. Subsequently, two categories of individuals—one category consisting of people from the private sector and another category consisting of people from the public sector—may be selected to act as sorters. If the popular belief that *nationalization leads to overall improvement* is true, the people of the public sector are expected to place the favourable statements (towards nationalization) high and the unfavourable statements low while sorting. The opposite expectation is due from people of the private sector. This type of study illustrates a one-way structured Q-sort because there is one basis of classification. To illustrate a two-way structured Q-sort, two variables—attitude and specificity—might be investigated. An attitude item may indicate either a liberal attitude or a nonliberal attitude, and at the same time it may be either a specific or general item. Some liberals might sort specific liberal items as high, whereas other liberals might sort general liberal items as high and so also the nonliberals. A schematic representation of a 2 × 2 factorial design thus constituted has been shown in Table 12.3.

Table 12.3 A two-way structured Q-sort with the number of items in each cell

	Specificity	
	Specific	General
Liberals	25	25
Nonliberals	25	25

The investigator might select 25 items which are both liberal and specific, 25 that are both liberal and general, and so on, for the bottom two cells of a 2 × 2 table. The above example illustrates the two-way structured Q-sort. Usually, the two-way structured Q-sort is very helpful in formulating a certain theory because the factorial design permits two variables to be related in a logical and empirical way (which is the essence of any theory).

The Q-methodology is controversial. As such, some have praised it and some have criticized it. Some of its merits on the basis of which the technique has been commended are as given below.

1. The chief merit of Q-sort is that it permits the subjects to make comparative responses towards a large number of stimuli. Ordinarily, it has been found that the reliability (usually the retest reliability) of the comparative responses obtained from Q-sort is higher than the reliability of the comparative analysis obtained from other absolute ratings. This is one of the important advantages of Q-sort.

2. Q-sort, particularly the two-way structured Q-sort, is very helpful in formulating a theory for testing the principles of theory because it allows the manipulation of variables in a logical and empirical way.

3. Q-sort data allow the operation of correlational methods as well as the analysis of variance. As such, the data facilitate statistical computation as well.

4. Since Q-sort provides a challenging and realistic task, the technique appears to be interesting to many.

The primary limitations of Q-sort are as given below.

1. Q-sort assumes without any reasonable proof that a normal distribution (or a distribution approaching normal) best fits the rating of each sorter.

2. Q-sort only provides an estimate of comparative ratings for each trait. It does not say anything about the strength of each trait within the individual.

3. Q-sort does not provide scope for the inclusion of a cross sectional or larger sample. Moreover, in this technique the investigator least cares for drawing a random sample of persons. Due to the above two reasons, the conclusion of Q-sort can be generalized to a larger population only with great risk.

4. Q-sort allows a forced-choice sorting which has several adverse repercussions. Many individuals feel irritated with the forced-choice ratings as it does not allow a free play of attitude, and at the same time requires them to conform to an unnatural, unrealistic and unreasonable requirement. In forced-choice Q-sorting, when the coefficient of correlation is computed (which is usually done), the information regarding the elevation (mean) and the scatter (standard deviation) of the set of scores is lost. Not only this, the forced-choice technique affects the sorting of the Q-cards, that is, the placement of one Q-card affects the placement of another Q-card. This automatically means that the Q-sort, being a forced-choice technique, violates the assumption of independence which underlies most statistical tests.

Despite these weaknesses, the Q-sort is a very useful research technique for problems, which require comparative reactions towards a larger set of stimuli.

Semantic Differential Scale

The semantic differential (SD) scale is an attitude measuring device developed by Osgood et al. (1957). The scale grew out of the studies of Osgood and his colleagues on the mediational theory of learning where the 'meaning' of stimulus is one of the central factors. While they were studying the structure of the domain of meaning represented by adjectives, they were in search of an instrument which could measure the various facets of meaning. The semantic differential scale is one such instrument. The semantic differential scale is similar to the Likert scale in that the respondent indicates an attitude or opinion between two extreme choices.

The SD scale may be defined as a collection of subscales (each anchored on the extremes by bipolar adjectives) in which absolute ratings of concepts are done. The SD scale is different from Q-sort in the sense that it does not permit a *comparative rating* of responses whereas Q-sort does. In the SD scale, the term 'concept' refers to the object which is to be rated. The purpose of SD is to measure the various facets of meaning of the concept, the various facets of meaning bein represented by adjectives. The reason why adjectives are selected to convey the meaning of th concept is that in our day-to-day language, most of our ideas are appropriately communicate only through adjectives.

For understanding the SD scale, it is essential to know the meaning of the term 'meaning Meaning is a very general term and it includes the various reactions of people towards an object There are three facets of meaning—denotation, connotation and association. *Denotation* refer to description of object(s) in terms of their physical properties. For example, an elephant may b described as being a big fat animal, mostly blackish, having a long trunk in the front, etc *Connotation* indicates the sentiments and feelings of persons about an object. It thus indicates the general implications that the object has for the person. The person may describe an elephant as 'a very dangerous animal'. *Association* refers to those objects which come to the mind of the person when he has heard or seen a particular word. Thus, a person might think of words like 'wild', 'jungle', 'tiger', 'lion', etc., after listening to the word 'elephant'. The semantic differential scale is a measure of mainly the *connotative* facet of meaning. Very few and limited evidences are available to show that even a *denotative* facet of meaning can be employed in a semantic differential scale.

In the semantic differential scale, the concept is usually rated on the seven-point scale having bipolar adjectives at the two extremes. Some have used a nine-point scale but a seven-point scale is more common and effective. For children, a five-point scale is sufficient. The first step in constructing the SD scale is to choose the concept. In doing so, two primary considerations are important. *First,* the concept must be that object or stimulus which can elicit different responses from individuals. If the concept is such that all people respond in the same way (that the variance in response among persons is zero or nearly so), the concept is not worth investigating. *Second,* the concept must be relevant to the problem being investigated. For studying the political atmosphere of a country, the choice of concepts like 'Student', 'Teacher', etc., would be examples of irrelevant concepts. However, for studying the educational environment these concepts would be most relevant.

When the concept has been selected, the next step is to choose the scales or adjective pairs. Again, there are two important considerations to be kept in view. *First,* it must be decided by the investigator as to what factor(s) should represent the adjective pairs. Osgood et al. (1957) have, on the basis of their factor analytic studies, discovered three kinds of factors, namely, Evaluation (E), Potency (P) and Activity (A) for the purpose. These factors are called *cluster of adjectives* and represent three dimensions of meaning along which a concept can be measured. These dimensions are technically known as *semantic* space. E factor is defined by pairs of adjectives like Good–Bad, Fair–Unfair, Clean–Dirty, Honest–Dishonest, and so on; P factor is defined by pairs of adjectives like Strong–Weak, Large–Small, Hard–Soft, Dominant–Submissive, and so on; and A factor is defined by pairs of adjectives like Hot–Cold, Active–Passive, Tense–Relaxed, Quick–Slow, and so on. Of these three factors, the E factor is the strongest because the pairs of adjectives of this factor have sharp bipolar extremes, that is, all pairs have very clear-cut positive and negative extremes. The investigator may wish to have scales of E factor or he may have scales of any two factors or include all the three factors in preparing the SD scales. In the study of attitude or value, it is wise to include only the scales of E factor. Besides these three factors, sometimes the investigator may wish to include scales of other factors like Tautness, Warmth, Familiarity, etc., but these are very uncommon factors. *Second,* the pairs of adjectives must be relevant to the concept being rated. For the concept 'War', pairs like Honest–Dishonest, Sweet–Sour, and so on would be irrelevant pairs but pairs like Heavy–Light, Quick–Slow, Hostile–Amiable and the like would be relevant pairs.

In the final preparation of the SD scale, each concept is written on a separate sheet of paper (preferably in the top-middle of the bipolar adjectives) with the same set of scales and the subject is asked to rate the concepts as he (or she) sees them. An example of the SD scale with 'Teacher' as a concept is given in Table 12.4.

Table 12.4 A sample of semantic-differential scale

TEACHER

1. Original	___: ___: ___: ___: ___: ___: ___:	Conventional	(P)
2. Reserved	___: ___: ___: ___: ___: ___: ___:	Outspoken	(E)
*3. Passive	___: ___: ___: ___: ___: ___: ___:	Active	(A)
*4. Slow	___: ___: ___: ___: ___: ___: ___:	Fast	(A)
5. Good	___: ___: ___: ___: ___: ___: ___:	Bad	(E)
*6. Shallow	___: ___: ___: ___: ___: ___: ___:	Deep	(P)
7. Fair	___: ___: ___: ___: ___: ___: ___:	Unfair	(E)
*8. Dull	___: ___: ___: ___: ___: ___: ___:	Sharp	(A)
*9. Warm	___: ___: ___: ___: ___: ___: ___:	Cold	(P)
10. Systematic	___: ___: ___: ___: ___: ___: ___:	Erratic	(P)
*11. Worthless	___: ___: ___: ___: ___: ___: ___:	Valuable	(E)
12. Successful	___: ___: ___: ___: ___: ___: ___:	Unsuccessful	(E)

*Meaning given in the text.

In Table 12.4, a sample of 12 scales of the concept 'Teacher' has been shown. Items bearing asterisk indicate that the pairs of adjectives have been reversed at random in order to reduce certain response biases like thoughtlessness and stereotyped reading. The letters in the brackets on the right indicate the factor to which the pairs belong.

Like Q-sort data, SD data can be analyzed for one individual only as well as for a group of individuals. The scores on the individual scales are first located and then summed up to find out the mean of the set of scores. Subsequently, the means are compared. If the two (or more than two) means are quite similar, it is indicated, then, that the two concepts are alike in meaning for individual(s); if the two means vary, it is indicated that the two concepts (which yielded those two means) are separated in semantic space. In such a case Osgood et al. (1957) have recommended the use of a special statistic called the *D statistic*, which is a measure of the linear distance between any two concepts. On the SD scale, a score means simply the assignment of the numeral checked by the subject.

Fair 7 : 6 : 5 : 4 : 3 : 2 : 1 Unfair

For example, if the person checks a concept on a pair of fair–unfair, say, in the third set of dots on the right, his score on this scale of fair–unfair would be 3. Similar procedures are adapted in scoring all the scales of the semantic differential.

The semantic differential scale has some *advantages* and *disadvantages*. The scale provides a very convenient and quick way of gathering impressions on one or more than one concept. Since the scores are summed over the different scales, they tend to average out the peculiarities, if any, among the scales as well as provide a basis for finer discriminations among the individuals. The most common disadvantage of the semantic differential scale is that the appropriateness of pairs of adjectives is questionable and little consensus exists among the experts regarding the suitability of the pairs selected. It is also said that often the responses given by the subjects are at a superficial and verbal level. For example, a person might value the concept 'Teacher' at high but might not really be attaching a higher value or meaning to the teacher. This criticism is, however, not unique to the SD scale; rather, it is common to all attitude-assessment techniques.

Behaviourally Anchored Rating Scale

These rating scales are relatively new ones and through them, the performance evaluation in a job or a profession is measured. The development of the scale is a lengthy process and is done in three stages. The first stage is one where a group of experts, usually through discussion, decides upon the possible dimensions, which can be used as the aspects of the performance being evaluated. Suppose, for example, the *academic performance* of a college is to be evaluated, and a group of experts consisting of students and teachers decides the following dimensions to be important and appropriate ones:

1. Sincerity of students
2. Observational ability of the teachers
3. Organizational ability of the teachers
4. Appropriate syllabus

Several statements regarding each dimension are prepared and this marks the beginning of the second stage. One example of the critical incident (or statement) of the dimension of observational ability is: when most of the pupils do not understand the lecture, the teachers should immediately sense this and try to find out where the defect lies. In this way, several statements regarding the different dimensions are prepared. Thus, the critical incidents are prepared in the form of small events to represent the superior, average and inferior performance of each dimension. Usually, the critical incidents are prepared by the experts. Subsequently, a group of raters is assigned to sort each critical incident to the different dimensions to which it belonged. On the basis of perfect or near-perfect agreement the incidents are finally selected and the incidents showing least agreement are discarded. The third stage consists of calculating the scale value and consistency of each critical incident. For this purpose, statements are given to a group of judges who rate all these statements on a scale of excellence. The median value of rating of each statement is calculated. The median value becomes the scale value for each statement. Finally, only those statements are retained that display higher agreement among judges (indicating a higher degree of consistency) and that yield a wide range of average scale values. In this way, the scale is ready.

Behaviourally, anchored rating scales have some advantages and disadvantages. Campbell et al. (1973) have shown that the behaviourally anchored rating scales, as compared to other rating scales, yield less halo error, less leniency error and fewer variances attributable to the methodology. The common disadvantages are that the procedure of the construction of the scale is time-consuming and lengthy, and the scale loses the trait of simplicity due to its methodological sophistications.

Nominating Technique—Sociometry

A nominating technique is a technique in which each person names or nominates other persons, events, objects, subject matters, which are perceived as fitting into certain categories or situations. The technique is commonly applied for studying social choices and rejections. Sociometry is one type of a nominating technique which is used in the study of group structure, social status and personality traits. The basic principles of sociometry were first enunciated by

Moreno in a volume entitled *Who Shall Survive?*, published in 1934. Sociometry may be broadly defined as a method of discovering and evaluating group structure, social status and personality traits through measuring the acceptance or rejection between individuals in a group. Thus it is a technique of evaluating interpersonal relationship in a group. Stanley & Hopkins (1978, 403) have defined sociometry as "the study of interrelationship among members of a group, that is, its social structure: how each individual is perceived by the group." With the help of the sociometric technique the data relating to the choice, communication and interaction patterns of individuals in groups are gathered and analyzed (Kerlinger 1973, 1986).

In studying the structure of the group through a sociometric test, each individual in a group is required to make one, two or three choices (sometimes, larger than three) for other persons for a specified purpose. For example, in a group of schoolchildren each child may be asked with whom he/she would like to sit or go to the circus or play, or who are the three best (or worst) pupils in class? In a factory, a worker may be asked to name the three individuals in order who command the maximum prestige in their group. Or, a worker may be asked to name two persons who have certain specified traits like restless–quiet, humour–humourless, and so on. An item in each such pair designating the favourable trait is given a positive score (say +1) and an item displaying the unfavourable trait is given a negative score (say −1). A person's score on each trait is the algebraic sum of positive and negative scores given by his peers in the group.

Data obtained on the basis of sociometric tests are usually analyzed by three principal methods: *sociometric matrix, sociogram* and *sociometric index*. Of these three, the first two are very popular and hence, will be given a wider coverage than the last one.

Sociometric matrix, also known as *sociomatrix*, is a simple cross-tabulation or rectangular array of $n \times n$ dimensions, n being equal to the number of individuals in the group. The meaning of the matrix can be illustrated through an example: Suppose a group of eleven students (A, B, C, D, E, F, G, H, I, J and K) in a class was put a sociometric question: "With which two members of this group would you like to go for a picnic?" If a member chooses the other member, it is displayed by 1; if he does not make a choice, it is shown through 0. The data (hypothetical) are presented in Table 12.5.

Table 12.5 Sociometric matrix of an 11-member group with two choices (11 × 11 matrix)

CHOSEN

	A	B	C	D	E	F	G	H	I	J	K
A	0	1	0	1	1	0	0	0	0	0	0
B	0	0	0	1	0	0	0	1	0	0	0
C	0	0	0	0	0	1	0	0	0	0	0
D	0	1	0	0	0	0	0	1	0	0	0
E	0	1	0	0	0	0	0	0	0	0	0
F	0	0	1	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	0	0
H	0	1	0	0	1	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	1	1
J	0	0	0	0	0	0	0	0	1	0	1
K	0	0	0	0	0	0	0	0	1	1	0
Sum	0	4	1	2	2	1	0	2	2	2	2

It is convenient to read the above matrix from left to right row-wise. The first row can be read as A where one does not choose C, F, G, H, I, J and K but chooses B, D and E. The matrix is usually analyzed by examining who chooses whom. Ordinarily, there are three kinds of choices—*simple,* *mutual* and *no choice*. A simple choice is one in which one person chooses the other person but he (the other person) does not choose him (the first person). Thus, the choice is one-way. For example, in Table 12.6, C chooses F but F does not choose C; G choose C but C does not choose G. Mutual choices are the two-way choices, that is, both persons choose each other. For example, in the above matrix B chooses D and D also chooses B; B chooses H and H also chooses B. The sum of the column of matrix indicates to what extent a particular member is chosen by other members of the group. We can say that the greater the sum, the higher the popularity. The matrix indicates that B receives 4 choices which means it is chosen by many members of the group. Thus, B is popular. A and G are not chosen by any one. D, E, H, I, J, and K each receives two choices, C and F receive one choice each. Obviously, C and F are not popular in the group. The technique of sociomatrix has the following *advantages*:

1. The relationship between each single pair is well recorded.
2. Joint relationships between all pairs are also recorded.
3. Two or more than two matrices can be easily combined and compared.
4. Sociomatrix is valuable in mathematical analysis and synthesis.

The sociometric matrix is, however, considered as an inferior graphic device to the sociogram (to be discussed later) in studying group structure. *Sociogram, or directed graph,* is another method for analyzing the data obtained from sociometric tests. Sociogram may be defined as a pictorial technique to produce a set of nominations or choices. Here, a simple or one-way choice is represented by a one-arrowed line like → and the mutual or two-way choice is represented by a two-arrowed line like ↔. A sociogram of the data presented in Table 12.5 is given in Figure 12.2.

An analysis of Figure 12.1 reveals several interesting points about the structure of the group. B is surrounded by arrows, which means he is chosen by several members of the group. A is a rejectee as he chooses B, D and E but neither of them choose him. A has no arrowheads pointing at him. I, J and K form a *clique*, which refers to that structure of the group in which three or more persons choose each other. The other common patterns are *mutual pairs* and *isolates*. Mutual pair refers to a pair in which the two members like each other but no other person in the group likes them. C and F are mutual pairs. Isolates are those who do not choose any one and are not chosen by anyone in the group. G is an isolate.

A comparative study of the sociometric matrix and the sociogram reveals that both tell the same story. The form of the sociomatrix is tabular whereas the form of the sociogram is pictorial. The obtained sum of choices in each column in the sociometric matrix is equal to the sum of

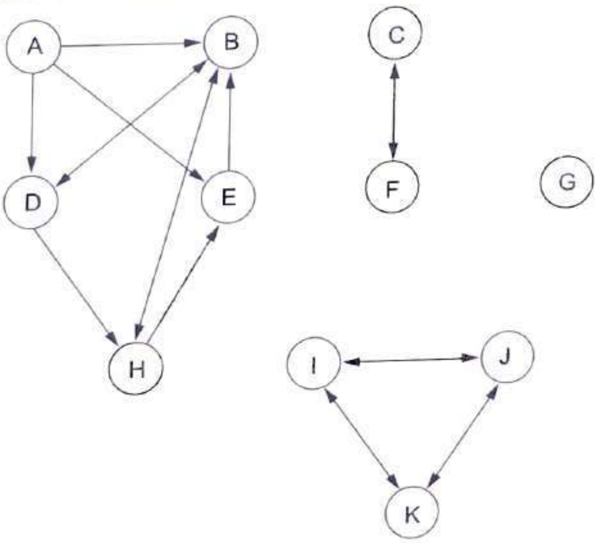

Fig. 12.2 A sociogram of the choices of eleven persons

arrowheads pointing at a particular person in the sociogram. B receives four choices in Table 12.6 and he has four arrowhead pointing at him. This is similar in the case of other persons in the sociogram, and in the sociomatrix. As a matter of fact, sociogram and sociomatrix supplement each other and in a research work, the two should be synthesized and used in a complementary fashion.

Sociograms have certain *limitations*. Sociograms tell the investigator only what persons are chosen or rejected and not the reason for rejection or acceptance. Sociograms revealing a certain group structure are only a tentative picture of the group. A sociogram prepared at the beginning of a certain class will be different from the sociogram prepared at the end of the session of the same class. Moreover, choices and rejections in the group are determined by several factors like sex, caste, prestige of the persons being chosen, rivalries, etc. Sociograms provide no way to check to what extent these factors are influencing the rejections or acceptances in the group. A person may not choose anybody for making friends nor be chosen by any other persons in the group, but this does not mean that he has no feeling of friendship at all in the group. Therefore, Thorndike & Hagen (1977, 484) have rightly concluded, "a sociogram is at best a rough and tentative picture of social currents and climate of the group."

Sociometric indices are the next important methods for analyzing sociometric data. One common index is the *choice status* of a person, which is given by Equation 12.1 as indicated below.

$$C_S = \frac{\Sigma C}{n-1} \tag{12.1}$$

where C_S = the choice status of the persons, ΣC = sum of choices a person receives and n is the number of persons in the group. Illustrating the calculation of C_S from the data of Table 12.6, B can be said to have $C_S = 4/(11-1) = 0.40$ and the C_S of E $= 2/(11-1) = 2/10 = 0.20$. The value of C_S becomes the direct index for the individual's popularity in the group.

Besides the above sociometric analyses, some sociometric instruments have been developed for evaluating the interpersonal relationship among pupils in the class. One such instrument is the Ohio Social Acceptance Scale in which each pupil rates every other pupil in one of six categories: (i) My very, very best friends; (ii) My other friends; (iii) Not friends, but okay; (iv) Don't know them; (v) Don't care for them; (vi) Dislike them.

On the basis of the ratings by all pupils, a sociometric score is computed for each pupil. The score indicates the extent of his acceptance in the group. The Syracuse Scales for Social Relations is another important sociometric instrument to assess the social relations is a class. The scale has three separate forms of which the first form is meant for the pupils of elementary classes (Grades 5 and 6), the second form is for pupils of junior high school (Grades 7 to 9) and the third form is for pupils of senior high school (Grades 10 to 12). Each pupil is asked to rate every other pupil on two important psychological needs in each of the three forms. In other words, each pupil rates every other classmate as a source of aid when he is in trouble with some personal problem (need-succorance), and each pupil also rates these classmates as a source of support in his effort to attain some personal goal, which may bring social commendation and approval (need-achievement). Results are usually summarized in the form of mean ratings received by a single pupil. The mean ratings indicate the extent to which each pupil feels favourable towards his classmates.

A sociometric matrix usually gives the following information about the group.

(i) In a group, there are a few persons who are liked by many. Such persons are very popular in the group. Such persons have the highest score at the bottom of the column of the score. In Table 12.5, B is such a student whose score is maximum, that is, 4.

(ii) In a group, there are persons who like others but are themselves not liked by others. Such persons are called rejected persons. In Table 12.5, A is a rejectee.

(iii) In a group, there are persons who form a clique of three persons in which all the three like one another but no other person of the group likes them. In Table 12.5, I, J and K form a clique. In a clique, the number of persons may be more than three.

(iv) In a group, sometimes a mutual pair (of two) is formed in which both the persons like each other. No third person likes them nor does any of the two like the other members of the group. In Table 12.5, C and F constitute an example of a mutual pair.

(v) In a group, there are persons who don't like any member of the group nor are themselves liked by any member of the group. Such persons are called isolates. In Table 12.5, G constitutes the example of an isolate.

Group sociometric measures are perhaps still more common. A measure of cohesiveness in a group through sociometric index can be completed by Equation 12.2.

$$C_o = \frac{\Sigma(i \leftrightarrow j)}{\frac{n(n-1)}{2}} \tag{12.2}$$

where, C_o = group cohesiveness

$\Sigma(i \leftrightarrow j)$ = sum of mutual choices (or mutual pairs)

n = number of persons in the group

In Table 12.5, $\Sigma(i \leftrightarrow j) = 6$, that is, there are six mutual pairs: C & F, I & J, J & K, I & K, B & D, and B & H. Thus, the group cohesiveness by Equation 12.2 will be

$$C_o = \frac{6}{\frac{11(11-1)}{2}} = \frac{6}{\frac{110}{2}} = \frac{6}{56} = 0.11$$

Obviously, the level of group cohesiveness in the group is low.

There is still another sociometric index which is used when the number of choices which each individual is permitted, is limited (such as, two choices or three choices or any fixed number of choices).

In such a situation, Equation 12.3 is applied:

$$C_o = \frac{\Sigma(i \leftrightarrow j)}{dn/2} \tag{12.3}$$

where, d = number of choices each individual is permitted; other symbols are defined like those in Equation 12.2.

Suppose each student of a group ($n = 5$) is asked to name two such students with whom he wants to play a game. Further, suppose that $\Sigma(i \leftrightarrow j)$ in that group = 4. Therefore, C_O by Equation 12.3 will be:

$$C_o = \frac{4}{\frac{2 \times 5}{2}} = 4/5 = 0.80$$

Obviously, the degree of cohesiveness in the group is high.

One advantage of sociometric test is that it is an easy and economical means of having information about the affected structure of the group. Another advantage is that such test possesses the trait of flexibility. As such, this can be used easily in different situations for different purposes.

Despite these advantages, sociometric test has some disadvantages. One great disadvantage is that such a test can be advantageously used only when the number of persons in the group is less than 20. When the number exceeds 20 or increases substantially, a sociometric test proves to be a cumbersome means for having information about the affected structure of the group.

PROBLEMS IN OBTAINING EFFECTIVE RATINGS

The factors which limit or affect sound and valid ratings may be conveniently divided into two parts: factors which limit the *raters' willingness* to rate the object or persons honestly in accordance with the given instruction, and factors which limit the *raters' ability* to rate accurately and consistently, even when the raters intend to do so. A detailed discussion of these two sets of factors is presented below.

Factors Affecting Raters' Willingness

In general, rating is an arduous and boring task. In order to obtain sound ratings, the investigator refines the instrument carefully by introducing elaborate procedures and precautions and gives very planned instructions to be followed by the raters in order to get rid of subjective impressions and superficial reactions of the raters. Theoretically, this is good and desirable but practice has revealed that some raters are not willing to take pains in rating the persons or objects in accordance with the set procedures and instructions. As a consequence, they do the ratings perfunctorily and hurriedly, keeping an eye more on finishing the task than upon making the accurate judgement. This has the natural consequence of lowering the effectiveness of ratings by bringing down their reliability and validity.

Raters' identification with the ratees may be another factor that limits the effectiveness of ratings. Often the rater is close to the persons being rated. He may identify with some persons positively and with some persons negatively. Persons identified positively are likely to be rated high despite some unfavourable characteristics, whereas persons identified negatively are likely to be rated low despite some favourable and sound characteristics. In both the situations, the rating would be a misleading index and would considerably lower the effectiveness of the ratings.

Thus the raters' willingness to rate honestly and conscientiously is limited by two factors: first, they may not be ready to take pains to do the rating according to set procedures and instructions; and second, they may identify with persons being rated either positively or negatively, which will colour their judgement.

Factors Affecting Raters' Ability

Sometimes raters are willing to make their ratings objective and accurate but some hindrances may distort their best intention to do effective ratings. Such hindrances directly affect the raters' ability to rate accurately. Thorndike & Hagen (1977) have mentioned the following factors, which are likely to affect the raters' ability to rate effectively.

1. *Opportunity to observe the individuals being rated*: For effective rating, close contact between the rater and the persons being rated is essential. If the persons are not in close contact with the rater, he may not be able to rate them with respect to the trait under study, even if he has the intentions to do so. For sound ratings, it is also essential that the rater be given a specific opportunity to make observations of the persons' behaviour. When a teacher is asked to rate his pupils on the trait of initiative and if he is rating only on the basis of his classroom observation, his ratings may not represent correct judgement. A pupil may take initiative in his classroom activities but the same pupil may not take initiative in a picnic, home and other social gatherings. Thus rating based simply upon general observation in the classroom may be a misleading index. So, general opportunity and specific opportunity to observe the individuals being rated are the factors which influence the effectiveness of ratings.

2. *Subjectivity in the traits being rated*: Some personality traits may be revealed directly in an interaction with other persons or events whereas some traits are not directly revealed, and the presence or absence of such traits is largely dependent upon the wise inferences and intentions of

the investigator. The former type of traits is the objective or overt traits of personality and the latter type of traits is *subjective or covert* traits of personality. Talkativness, dominance, outgoingness, punctuality, etc., are some examples of overt traits whereas feelings of insecurity, anxiety, self-satisfaction, ego-strength, emotional control are some of the covert traits of personality. In general, the covert traits are rated with lower reliability and validity than the overt traits. The reason is that the rating of covert traits is largely dependent upon the inferences made by the raters and they are least involved in outward interaction with other persons. Since very little regarding such traits comes to the surface, they are rated with greater difficulty and inaccuracies. Thus, the covertness of traits is a factor which lowers the reliability and validity of the ratings.

3. *Vagueness in the meaning of the trait rated*: Some traits or dimensions to be rated are vague and abstract ones. As a consequence, their meaning varies from rater to rater and this naturally affects the consistency in ratings. On the other hand, when traits or dimensions are well defined, broad and have a pin-pointed reference, there occur less variations in their meaning from rater to rater. As such, their ratings are more consistent. Talkativeness is one such well-defined trait and it is very likely that all the raters will mean the same thing when they use this term. When a group of teachers is asked to rate the pupils in a class on the trait of being talkative, variations are less likely to occur in ratings by the teachers because all of them would mean the same thing. On the other hand, when the same teachers are asked to rate the same pupils on the trait of self-sufficiency (an example of a vague dimension), the ratings are likely to lose consistency (or reliability) because teacher A may use this term to mean that a student has all things necessary for the development of his self; teacher B may mean that a student has self-respect, self-control and self-satisfaction; and teacher C may mean simply that a student has sufficient physical growth to ensure proper development of his self or personality. Thus vagueness or ambiguity in meaning of the trait or dimension being rated is likely to influence the reliability of the ratings.

4. *Nonuniform standard of reference*: Most of the rating scales require the raters to rate the ratees in any one of different categories:

Superior, excellent, very good, good
Satisfactory, unsatisfactory
Best, good, average, fair, poor
Good, sufficient, average, below average, less

When a rater rates a ratee as superior, the question arises as to what the standard is against which a ratee is being classified as superior. Is he being compared with the top 5% or top 10%? Or, is he being compared with the middle cases? The raters have no uniform standard before them so that the interpretation of category 'superior' may not have an identical meaning to all. In the absence of a uniform standard, the interpretation of a category varies from rater to rater. One rater's 'superior' may be another rater's simply 'very good'. This naturally lowers the consistency of ratings and thereby, their reliability.

5. *Raters' personal characteristics*: Raters' personal characteristics also tend to influence the ratings. Some raters are conservative and, therefore, they tend to rate persons almost in the middle. They rarely rate a person very high or very low. Some raters are tough and, therefore, rarely rate anybody high. These are popularly known as errors in ratings which due to their special importance, have been separately explained later. Besides these general characteristics, some personal experiences also tend to affect the ratings. A teacher who is annoyed with the behaviour of a student in a classroom is likely to rate him low; a supervisor in an industry may be pleased with the way a worker behaves with him and therefore, is likely to rate him high, whether his overall performance is satisfactory or not. An officer may be annoyed with his young female stenographer and typist simply because she ignores him. In such a situation, she is likely to be rated low, no matter how good her performance may be. Thus personal experiences of raters are

likely to colour their judgement and also their ratings. In such situations, the ratings done by one rater cannot be compared with the ratings done by another rater. Personal experiences, therefore, are the potent factors, which influence the ratings.

In view of the above limitations, it is suggested that rating scales can be an effective means of appraising others only when the points given below are given due weightage.

METHODS OF IMPROVING EFFECTIVENESS OF RATING SCALES

Psychologists have devised some means to improve the effectiveness of such rating scales. Given below are the important means.

Refinement in Stimulus Variables of the Rating Scales

In order to improve the effectiveness of rating scales, it is essential to make refinements in *stimulus variables* of the rating scales. By *stimulus variable* (S) is meant the traits, dimensions or qualities to be rated. As stated earlier, some traits are objective and easily observable and, therefore, can be rated with greater accuracy and consistency than traits which are subjective and ill-defined. Researches on the different types of traits appropriate for higher accuracy, consistency and objectivity in ratings have revealed that traits like punctuality, intelligence, friendliness, clearness, leadership, physical health, scholarship and independence are rated with higher consistency and accuracy whereas traits like courage, co-operativeness, depression, popularity, sympathy, kindliness are rated with wider disagreements among raters. From these various researches some guidelines regarding the traits have been prepared which, if followed, are likely to produce higher agreement and consistency in ratings. These guidelines are given as follows.

1. Each trait should clearly refer to only one type of activity.
2. A trait should not represent a combination of a number of traits that vary independently.
3. Traits should be such as can be defined in objective, unequivocal and specific words.
4. Traits should be judged on the basis of past and present accomplishment rather than on the basis of future accomplishment.
5. In defining a trait words like 'often', 'always', 'very', 'extremely' should be avoided because these are general words and give no hints regarding the appropriateness.

Refinement in Response Variables of the Rating Scales

The response variable of the rating scales refers to the response option, which consists of numerical or adjectival categories. As mentioned earlier, these categories in terms of which the rating is done, are very subjective. One person's 'superior' is another person's 'outstanding'. Not only this, these categories also allow the free play of a rater's bias. It is, therefore, argued that refinements in response categories are essential so that the objective interpretation of categories and therefore, the objectivity in ratings can be achieved. One of the most fruitful attempts in this direction has been to develop a rating form that forces the raters to discriminate between two or more than two equally desirable or undesirable response options. The second attempt is to develop a rating form which controls raters' differences in judging standards. An example of the first attempt is the forced-choice rating pattern and examples of the second attempt are percentage scale, graphic scale, behaviourally anchored scale, man-to-man scales, etc. All these techniques have already been discussed in detail.

Improvement in Rating Procedures

Even when the stimulus variable and the response option of the rating scale have been refined, the ratings cannot be objective and consistent unless their procedures are improved. There are several factors that go together in improving the rating procedures. These factors are given below.

1. *Selection of raters*: The selection of proper and appropriate raters is one of the fundamental aspects of rating procedures. A *proper* and *ideal* rater is a person who has been in

close contact with the ratees and has several opportunities to observe them in situations relevant to the trait on which they are to be rated. If the rater does not know the ratees intimately, he cannot rate them well, no matter how much he is motivated to do so. Sometimes, it may be necessary for selected raters to have a certain amount of training in the type of judgement they are required to make and in the use of rating instruments. Researches conducted in industrial as well as military settings have revealed that training which defines aims and purposes of the ratings and illustrates the meaning of traits to be rated, tends to increase the validity of ratings by decreasing constant errors like the halo effect, error of leniency and error of central tendencies. Besides these considerations, some other factors related to raters should also be given proper attention for ensuring effective ratings. For example, raters should be allowed sufficient time in their rating and the raters' educational and professional backgrounds should be similar to the ratees' educational and professional background.

2. *Improving the reliability of ratings*: One general limitation of the rating scales is that its between-raters reliability is very low. That is, if two or more than two independent raters have rated the same persons, correlations (a measure of consistency) between the raters are generally very poor indicating a low reliability. To improve the reliability of the ratings, such ratings done by independent raters may be pooled together provided all the independent raters are equally qualified and/or know the ratees equally well. Pooling of the ratings functions in the same way as increasing the length of the test and as reliability of a test is increased after its length is increased, so also reliability of ratings is increased after independent ratings are pooled. Like a lengthened test, the Spearman–Brown formula (Equation 5.17) may be used to estimate the reliability of pooled independent ratings. To illustrate, suppose the reliability of one rater is 0.45. Then the reliability of two raters and three raters by substituting in Equation 5.17, would be:

$$r_{nn} = \frac{2(0.45)}{1+(2-1)0.45} = 0.62 \quad \text{(For two pooled ratings)}$$

$$r_{nn} = \frac{3(0.45)}{1+(3-1)0.45} = 0.71 \quad \text{(For three pooled ratings)}$$

But in many practical situations, the pooling of independent ratings is not feasible because the rater may not be equally qualified and/or equally know all the ratees. For example, in a primary school each class has only one class teacher who knows all the pupils of his or her class more or less equally well. Other teachers may not know the pupils of the class to that extent. In such a situation, if the independent ratings of a particular class done by a group of several teachers (including the class teacher) are pooled, the reliability of ratings is likely to decrease rather than increase. A similar situation may occur in the case of rating workers in an industry or clerks in an office. Only the supervisor who is the officer-in-charge of the workers will know them well and the immediate officer who is in charge of the group of clerks will know them well. Other supervisors or officers may not be knowing them well and hence, the independent ratings by a group of supervisors or officers should not be pooled to estimate the reliability of the ratings. If it is done, chances are that the reliability will go down.

ERRORS IN RATINGS

Ratings are subject to varieties of errors, some of which have been referred to earlier. One general class of errors called *constant errors* (because they tend to throw the ratings constantly in one direction) consists of *halo effect, error of severity, error of leniency* and *error of central tendency*. Besides these constant errors, there are contrast errors, logical errors and proximity errors in ratings. A detailed discussion of each of these errors is presented below.

Halo Effect

Halo effect is one of the very common constant errors in rating. The error was first mentioned by Wells in 1907 and was named so by Thorndike (1920). In the words of Anastasi (1968, 420), halo effect refers to "a tendency on the part of the raters to be unduly influenced by a single favourable or unfavourable trait, which colours their judgement of the individual's other traits". Thus halo effect is produced when the rater rates the ratees in a constant direction of general impression, which he has formed earlier. The common example of the halo effect is the tendency on the part of a teacher to rate the quality of answerbooks of a pupil whom he likes as higher than the answerbooks of a pupil whom he does not like. Likewise, a teacher who is very much influenced by the outstanding performance of a student in a game, may rate his answerbook higher, though the quality of answer may not be up to the standard. Symonds (1925) has reviewed a series of experiments conducted to locate situations in which the halo effect is most obvious and has concluded that it occurs mostly in a trait which is not clearly defined, not easily observable, difficult to single out frequently, and involves moral importance or belongs to the traits of character.

Halo effect has two marked effects upon ratings. One obvious effect is to throw the ratings in either a positive or negative direction depending upon the general impression, and to that extent, the rating loses reliability and validity. The second effect of halo is to make the positive correlations between traits being rated spuriously high. Halo effect, therefore, must be controlled, though it is very difficult to avoid. One general way to reduce the halo effect is to define the trait in observable units of behaviour. Halo effect may also be counteracted when the rater rates all ratees on one trait before going on to the next trait. Thus when several ratees are to be rated by the same rater on several traits, the practice of rating all the ratees on one trait at a time forces the raters to pay attention to each trait separately rather than to the general impression of each ratee. Halo effect may be reduced by arranging the scale points in such a way that the desirable end of some traits fall on the right-hand side and the desirable end of some other traits fall on the left-hand side. This arrangement automatically prevents the rater to rate a ratee on the right-hand side for his general desirable impression and to rate him on the left-hand side for his overall undesirable impression. Forced-choice formulas may also be adopted to reduce the halo effect in ratings.

Error of Severity

The error of severity refers to a constant tendency (irrespective of the trait being rated) on the part of the rater to rate the ratees too low. Such persons are 'hard raters' and are usually reluctant to rate any person or object high. Kerlinger (1973, 549) defines the error of severity as 'a general tendency to rate all individuals too low on all characteristics.' Forced-choice type of rating, where all options appear equally attractive and where the rater is unable to throw the ratings low at his will, is one convenient technique for reducing the error of severity.

Error of Leniency

As its name implies, the error of leniency refers to a constant tendency, irrespective of the traits being rated, to rate the ratees too high. Such persons are 'easy raters' and try to concentrate on the higher end of the scale. Such raters are reluctant to rate anybody unfavourably. Like the error of severity, this error can be controlled by the forced-choice technique.

Error of Central Tendency

Error of central tendency refers to a constant tendency on the part of the rater to avoid extreme ratings and place his ratings in the middle or average category. Kerlinger (1973, 549) defines the error of central tendency as 'the general tendency to avoid all extreme judgements and rate right down the middle of a rating scale.' When the rater is asked to rate the individuals who are not well known to him he hesitates to rate them on the extremes and, therefore, prefers to rate them in the middle. In order to counteract this error, Guilford (1954) suggests that greater variation in meaning between the steps of the scale at the extreme than between the steps at the middle should be introduced.

Besides these constant errors, other types of errors common to rating scales are given below.

Contrast Error

Contrast error, first pointed out by Murray (1938), refers to a tendency on the part of the raters to rate persons or objects in a direction which is in contrast to the trait he himself possesses. Thus, through this error the rater projects his own attitude or bias in the ratees. For example, a teacher who is very punctual in coming to the class, may rate other students to be less punctual than he is. Similarly, a supervisor who possesses the trait of orderliness may rate others in a direction opposite to this trait, that is, he may rate others to be less orderly.

Proximity Error

Proximity error, first pointed out by Stockford & Bissell in 1943, is one which occurs due to proximity or nearness of two traits being rated. The nearness between two or more traits may be in terms of meaning or in terms of time. Co-operativeness and friendliness are more or less two similar traits in meaning. When the rater is asked to rate such traits without intervention by some dissimilar traits, the correlation between the ratings tends to be spuriously high. However, when these traits are separated by dissimilar traits, the correlations between adjacent traits and remote traits are nearly equal indicating the cancellation of proximity error. Thus proximity error tends to make the intercorrelations between adjacent traits (which are similar) higher than those of remote traits, though their degree of actual association may be equal. Proximity error may also be counteracted by allowing a sufficient time interval between the rating of one trait and another.

Logical Error

Logical error is similar to proximity error and halo effect in the sense that it also tends to increase the intercorrelation of the ratings of the two or more traits. Logical error, pointed out by Newcomb (1931), is one where the rater gives more or less a similar rating to traits which seem to him, somehow logically related with each other. Thus the logical presupposition or similarity (about the traits to be rated) in the mind of raters is the basis of the error. The logical error is most obvious where the traits to be rated are abstract and semantically overlapping. This error can be reduced, at least partially, when the raters are given the most objective and clearly defined traits for ratings.

EVALUATION OF RATING SCALES

Despite their limitations, rating scales undoubtedly continue to be one of the most promising appraisal techniques. Rating scales have a wider range of application. They can be used even by those persons who are not trained. They can be employed when a larger number of stimuli are to be rated and in this task, the rating scales yield far more consistent results than their rival, the ranking method or the method of rank order. The common observation is that when the number of stimuli to be ranked is large, say for example 80 or 90, the ranking becomes a difficult task and not only that, even the person who is doing the ranking, gets bored. Rating scales, and particularly the graphic scale, present a very interesting task for the rater and he seldom gets bored even with a long list of traits to be rated. Some comparative studies of ranking method and rating scales have revealed very significant facts. Conklin & Sutherland (1923) studied different kinds of jokes by the ranking method and the rating scales and found that the rating method yielded less variance from one expert to another than the ranking method. Symonds (1925) has reported that ordinarily the results yielded by rating scales are in no way less reliable than the results yielded by the ranking method. So far as the various errors involved in ratings are concerned, they are such that they can be controlled and corrected, if proper care is taken. Therefore, the rating scales, on the whole, have a bright future as an appraisal technique of traits, objects, events, etc.

MEANING AND FEATURES OF SECONDARY DATA

By secondary data we mean those data that are in actual existence in accessible records, having been already collected and treated statistically by the persons maintaining the records. In simple words, secondary data are data that have already been collected, presented, tabulated, treated with necessary statistical techniques and conclusions have been drawn. Therefore, collecting secondary data never means doing some original enumeration, rather it simply means obtaining data that have already been collected by others including government departments. When once primary data have been collected originally by someone else, it becomes secondary data in the hands of all other persons who are desirous of handling it for their own purposes. If Mr A collects some data originally then data are primary data to Mr A but the same data when used by Mr B, becomes secondary data to Mr B.

The major features of secondary data are as under:

(i) Secondary data are *ready-made* and in this sense, they are time saving. The researcher easily saves the time and labour involved in conducting interview or administering questionnaire.

(ii) The form and content of secondary data are shaped by others. Clearly, this feature of the secondary data can limit the overall scientific value.

(iii) Secondary data are not limited in time and space (Festinger & Katz 1953). It means that the researcher using secondary data need not have been present either *when* or *where* they were collected.

Keeping these features in view, it can broadly be said that the secondary data are items of information originally collected for a purpose other than its present scientific use.

Sources of Secondary Data

There are many sources of secondary data. Some of the important ones are as under.

1. *Private sources of secondary data:* Private sources of secondary data are strictly personal documents like letters, diaries, and other bibliographical materials such as individual life histories. These are considered very useful in examining social conduct. Another private source of secondary data exists in the files and published materials of private organizations that characterize most of the societies in the world.

2. *Public sources of secondary data:* Public sources of secondary data are those sources that rely on the information available in various data archives and other related sources. The following are some such important sources:

(i) Central and state government publications
(ii) Publications done by international organization like UNO, UNESCO, WHO, etc.
(iii) Official publications as well as reports of municipalities, zila parishads, etc.
(iv) Foreign government publications
(v) Reports and publications of various commissions like the UGC, ICSSR, trade associations, banks, stock exchanges, etc.
(vi) Well-known newspapers and journals
(vii) Internet and various website sources
(viii) Publications brought out by Supreme Courts, High Courts and Lower Courts

The above list of secondary data, however, can't be said to be thorough or complete because there are a number of other sources also such as manuscripts of eminent scholars, statisticians, labour bureaus, records of business firms, etc.

Verification of Secondary Data

It is suggested that before accepting secondary data, it is always essential to scrutinize them properly in regard to its accuracy and reliability. Therefore, verification of secondary data is essential and it should be carried out in the following manner:

(a) Whether the data was collected at the proper time

(b) Whether the organization that has collected the data is reliable

(c) Whether the data has timely relevance and is not outdated for the present research

(d) Whether the appropriate statistical tools, if any, were used by the primary data investigators

(e) Whether enumerators of facts and figures had any prejudice, bias, or showed negligence, hurry or carelessness in their compilation work

(f) Whether the data enumerators had ensured proper standards of accuracy. If so, up to what degree such accuracy has been shown.

Major Uses of Secondary Data

There are some major uses of secondary data. These various uses can be examined from two angles.

(i) Secondary data can be viewed from the standpoint of how they can really be used to generate scientific information.

(ii) Secondary data can also be viewed in terms of ways they promote the overall objectives of the scientific enterprise.

For those social scientists who support the viewpoint of the first use, there are three ways to proceed ahead.

(a) Such secondary data can be used to demonstrate original scientific meanings. Here we get good examples in Hyman's (1972) review of the works of the secondary data researchers.

(b) Such secondary data may serve as auxiliary documentation, that is, they can complement original data being gathered by the investigator.

(c) Secondary data can be used for verifying the previously established findings. In other words, when secondary data are appropriately used, they can provide a relatively crude form of replication.

So far as the uses of secondary data in terms of overall objectives of the scientific enterprise is concerned, all the above-mentioned applications are considered important.

In a nutshell, it can be said that whatever uses secondary data sources have for scientific activity, it depends upon the subjective ingenuity of the investigator in making use of them.

Advantages and Disadvantages of Secondary Data Analysis

The major advantages of secondary data analysis are as under:

(i) Such analysis saves time and money (Useem 1973).

(ii) Such analysis provides a good and real possibility of using the work of others to broaden the base from which scientific generalizations can be easily made. This is especially so when information from several cultural settings is being studied.

(iii) Secondary data analysis provides a ground for using such data in verifying the findings already obtained in primary data by an investigator. In fact, by turning to the sources already containing relevant information, the verification process is rapidly enhanced.

The major disadvantages of secondary data analysis are as under:

(i) Even under most optimal conditions, secondary data analysis poses difficult problems for the researchers. Those who hope to save time and reduce the amount of tedium connected with procedural details such as coding and classifying may become disappointed due to the lack of accuracy and reliability.

(ii) Another problem with secondary data analysis is that it is sometimes not possible to comprehend the process of original data getting. Simply because data are available does not mean that reasons are given for why they were gathered.

(iii) In secondary data analysis, the knowledge of the whereabouts of the sources is not necessarily available to all social scientists on an equal basis. Any one who is familiar with social organization of scientific research knows that accessibility depends to a degree on proximity. Being at distances from data archives can do much to hinder the potential accessibility of secondary information. This, in turn, can affect the knowledge an investigator might have of the types of sources.

(iv) In secondary data analysis, the investigator has to depend upon data already collected for other purposes. Such data must be classified in ways consistent with the objectives of the present research. At this stage, several technical problems do arise. The researcher may find the relevant variables are missing or not measured in ways that render them useful as data. He may also find that the data as they exist, don't conform to the categories and classifications envisioned by him. Due to these reasons, serious doubts about the overall worth of the secondary data source arise and many researchers may find it pleasant to abandon their projects at this point.

Thus, we see that although secondary data analysis proves very easy and time-saving, is not without its cost.

☑ Review Questions

1. What do you mean by a questionnaire? Discuss the characteristics of a good questionnaire.
2. Discuss the relative advantages and disadvantages of different types of questionnaires.
3. Make a distinction between questionnaire, opinionnaire and schedule. Discuss the major functions of a questionnaire.
4. Discuss the important types of observations. Make distinction between participant observation and nonparticipant observation.
5. Discuss the methods of content analysis. Point out the important limitations of content analysis.
6. Describe briefly the two main approaches to objective observation and elucidate the factors affecting the reliability of observational data.
7. Discuss the usefulness and limitations of sociometry as a tool of psychological research.
8. Discuss the major problems in obtaining effective ratings. How can they be resolved?
9. Discuss the important sources of errors in ratings.
10. Discuss the major functions of interview. Also, point out the major sources of error in interview.

11. Discuss the relative advantages and limitations of different types of interviews as used behavioural sciences.
12. Bring out the utility of semantic differential as a tool of psychological research.
13. What is meant by rating scale? Discuss the importance of different types of rating scales
14. Discuss the various methods of improving the effectiveness of rating scale.
15. Discuss the important types of errors involved in rating scale.
16. Discuss the major sources of secondary data. Also, point out the major advantages an disadvantages of secondary data.

◆

13

SCALING TECHNIQUES

CHAPTER PREVIEW

- Distinction between Psychophysical and Psychological Scaling Methods
- Psychophysical Scaling Methods
    - Method of Limits
    - Method of Constant Stimuli
    - Method of Average Error
- Weber's Law and Fechner's Law
- Stevens' Power Law
- Newer Psychophysical Methods
    - Method of Category Scaling
    - Method of Magnitude Estimation
- Psychological Scaling Methods
    - Method of Rank Order
    - Method of Successive Categories
    - Method of Pair Comparisons
- Methods of Attitude Scales or Opinionnaires
    - Method of Equal-Appearing Intervals
    - Method of Summated Ratings
    - Guttman's Scale, or Cumulative Scale

DISTINCTION BETWEEN PSYCHOPHYSICAL AND PSYCHOLOGICAL SCALING METHODS

Scaling methods are methods through which stimuli or individuals are sorted according to some known and specified characteristics or attributes. When these methods are successfully applied for sorting purposes, their results automatically yield a continuum which is called a scale. Depending upon the classes of stimuli (to which the scaling methods are applied), scaling methods can be grouped into two general classes—*psychophysical scaling methods* and *psychological scaling methods*.

In psychophysical scaling methods, the stimuli presented have a physical referent. In other words, the stimuli presented can be described by a physical scale e.g., grams, decibels, centimetres, minutes, and so on. The obvious purpose of psychophysical scaling methods is to investigate the quantitative relationship between subjective measurement of these stimuli (as judged by the subject) and the objective measurement by the physical scales. In other words, the psychophysical scaling methods intend to discover some definite quantitative relation between the physical stimulus and the resulting sensation (or experience) by manipulations of the physical stimulus dimensions.

The psychological scaling methods differ from the psychophysical scaling methods in that the stimuli they measure are such that cannot be described by a known physical scale. In other words, the stimuli here possess psychological attributes, which obviously lack an appropriate physical referent. In the preface to his famous book, *Theory and Methods of Scaling*, Torgerson (1958) has said, "Psychological scaling methods are procedures for constructing scales for the measurement of psychological attributes." For measuring psychological attributes like sense of humour, quality of beauty, extent of honesty, aggressiveness, pleasantness, and so on, there are no physical scales and, therefore, these attributes are best measured by the psychological scales, which give accuracy to the measurement of stimuli through the judgemental processes of individuals.

There are different types of psychophysical scaling methods and psychological scaling methods. Each of these types will be examined in separate sections.

PSYCHOPHYSICAL SCALING METHODS

Suppose that 20 objects of the same size and shape but of different weights are to be arranged in an order from the heaviest to the lightest. The obvious way to do this is to weigh them, record the weight of each object and arrange them in terms of their measured weights. Further, suppose that no scale is available for weighing the objects and they still have to be arranged in an order from the heaviest to the lightest. One obvious way to do this is to give all these objects to a group of individuals with the request to arrange them in order of heaviness. Still another way may be to present all the objects in all the possible pairs and ask each individual to judge which one of each pair is heavier. Subsequently, on the basis of the average judgements, all the 20 objects can be ordered from the heaviest to the lightest. The scale used for weighing the objects is known as the *physical scale* and the order in which the objects are arranged in terms of the measured weights is known as the *physical continuum*. The order in which the objects are arranged on the basis of judgements made by the subjects regarding the weight is known as the *psychological continuum*. Several attempts have been made to investigate the relationship between the ordering of objects on a known physical continuum and the ordering of the same objects on a psychological continuum formed by judgements of individuals. The methods used in studying such a relationship are known as *psychophysical scaling methods*.

The term 'psychophysics' owes its origin and name to G T Fechner (1801–1887) who defined it as "an exact science of the functional relations of dependency between body and mind." For the first time, he set out to explore the quantitative relationship between the magnitude of sensation occurring in the mind and the magnitude of the physical stimulus that produces the sensation. For investigating the quantitative relationship between the magnitude of sensation and the magnitude of physical stimulus, he developed some experimental methods, which are still in use today. Before examining these methods, it is essential to explain the meaning of some common terms used in psychophysical measurements such as *threshold, point of subjective equality, point of objective equality, etc.*

The concept of threshold was first introduced by Johann Herbart in 1824 when he defined the 'threshold of consciousness'. The Latin equivalent of the term 'threshold' is *limen*. Threshold refers to that boundary value on a stimulus dimension, which separates the stimulus that produces a response from the stimulus that makes no response or a different response. The threshold in psychophysical measurement is ordinarily divided into absolute threshold and difference threshold. *Absolute threshold, or stimulus threshold* (abbreviated to *RL* from its German equivalent *Reiz Limen*), refers to that minimal stimulus value which produces a response 50% of the time. A physical stimulus value which is below that minimal value fails to elicit a response. Andreas (1960, 99) has defined absolute threshold as "a boundary point in sensation, separating sensory experience from no such experience when physical stimulus values reach a particular point." Thus absolute threshold defines the minimum limit for responding to stimulation. *RL* for a single physical stimulus is not the same for different individuals. It varies from individual to individual, and sometimes from one situation to another for the same individual. Hence, some people may perceive a stimulus at a low value while others may perceive the same stimulus at a higher level. That is why *RL* is statistically defined as the mean of RL_s taken over several trials by the same subject for the same stimulus. The *difference threshold*, or *differential threshold* (abbreviated to *DL* from its German equivalent *Differenz Limen*), is the difference between the two stimuli, which can be perceived 50% of the time. Thus *DL* defines the individual's capacity to respond to difference in sensitivity. The difference threshold is also sometimes referred to as a *just noticeable difference* (abbreviated to *JND*) which is the smallest difference between the two stimuli that can be detected by the subject. A stimulus must be increased or decreased by one *JND* in order that the change be perceived. Usually, for calculating *DL*, two stimuli are presented to the subject. One of them has a constant value throughout its presentation and is known as the *standard stimulus* (S_t) and the second stimulus is varied throughout its presentation and is known as the comparable stimulus (C_o). Suppose the experimenter takes an S_t of 100-g weight and starts presenting C_o at a very low value (say 30 g). He may then go on increasing the value of C_o by a very small value (say at the rate of 5 g in each presentation) until it becomes indistinguishable from the S_t. This is called the *lower difference threshold*. Likewise, the experimenter may start presenting the C_o at a much higher value than S_t (say, 140 g) and then go on decreasing the C_o by a very small value (say by 5 g) until it can no longer be distinguished from S_t. This is known as the *upper difference threshold*. The upper and lower thresholds represent the upper and lower limits of the *interval of uncertainty* or *IU* respectively. The *IU* indicates the span or range where the responses of the subjects are uncertain. The *DL* or *JND* is half the difference between the upper and lower threshold, that is, *DL* = Upper threshold – Lower threshold/2 or *IU*/2. The *DL*, like the *RL*, taken over several trials by the same subject for the same physical stimulus is not identical in different situations.

RL, thus, is a point on a physical stimulus, which is detected 50% of the time whereas *DL* is not a point; rather, it is a span or distance or range where the amount of change in magnitude of the stimuli can be detected 50% of the time.

The point of subjective equality (*PSE*) is another important concept used in psychophysical measurement. It is defined as that value of C_o which is, on the average, judged by the subject to be equal to the value of S_t. Surprisingly, PSE is rarely equal or identical to S_t. It is a very useful concept in the measurement of the extent of illusion. In the Muller–Lyer illusion, the extent of illusion (or the constant error) is defined as the difference between the PSE and the point of objective equality (*POE*). The *POE* is defined as the exact value of S_t . Thus, the constant error (*CE*) which indicates the extent of the Muller–Lyer illusion is simply the difference between *PSE* and S_t and may be expressed by the following formula (Woodworth & Scholsberg 1954):

$$CE = PSE - S_t \tag{13.1}$$

Thus, when judgements or discriminations differ significantly from the standard stimulus, the presence of the constant error is assumed. If the *CE* is *negative* it indicates underjudgement and if it is *positive*, it indicates overjudgement. Thus, if a 50-mm line is used (arrow-headed line), and the mean of the subject's setting of the C_o (the feather-headed line) in the Muller–Lyer illusion is 45.5 mm, the *CE*, and therefore, the extent of illusion, is 45.5 mm – 50 mm = –4.5 mm.

Because *CE* is negative, it indicates underjudgement, that is, the feather-headed line appea longer than the arrow-headed line. In other words, the real judgement of arrow-headed line underestimated (at 45.5 mm). Thus the feather-headed line of 45.5 mm appears to be of 50 m providing evidence for being perceived as longer than the arrow-headed line. It should be clearl held that the constant error is not really constant from trial to trial. The word 'constant' does n mean that the amount of the error is equal in each successive trial or that the direction of err (positive or negative) is consistent over the successive trials for any given subject. Frequently, systematic tendency on the part of the subject to underjudge or overjudge a given comparabl stimulus with some standard stimulus is found. Such a systematic tendency is known as th *constant error.* Why is then the amount of error not equal in each trial? What causes fluctuatio from trial to trial? This is probably because the cause is not known with certainty. It may be due t the variations in motivation or attention of the subject or due to the unco-operativeness of th subject. Error produced by these factors and the like is known as *variable error* or *chance erro* Such errors work in both directions and together they tend to cancel each other out. Usually, in variable error the source of error is not known with certainty and does not throw the response in one constant direction. The term 'error' here does not indicate that one judgement is right and another is wrong. It simply indicates the extent of fluctuation in judgements from trial to trial. The variable error is indexed through the standard deviation (*SD*) of those responses judged to be equal to S_t. However, three factors, namely, subjects' sensitivity which may vary from moment to moment, slight unavoidable changes in the physical characteristics of the stimulus and changes in interest and attitude seem to be an important source of variable error.

Basically, there are three classical psychophysical methods which are frequently used. A detailed discussion of each is presented separately.

Method of Limits (also known as Method of Just Noticeable Difference, Method of Minimal Changes or Method of Serial Exploration)

The method of limits is a popular method of determining the threshold. The method was so named by Kraepelin in 1891 because a series of stimulus ends when the subject has reached that limit where he changes his judgement. For computing threshold by this method, two modes of presenting stimulus are usually adopted—the increasing mode and the decreasing mode. The increasing mode is called the *ascending series* and the decreasing mode is called the *descending series.* For computing *DL*, the C_o is varied in possible small steps in the ascending and descending series and the subject is required to say in each step whether the C_o is smaller (–), equal to (=), or larger (+) than the S_t. The data (hypothetical) given in the Table 13.2 illustrates the computation of *DL* by the method of limits. For computing *RL*, no S_t is needed and the subject simply reports whether or not he has detected change in the stimulus presented in the ascending and descending series. The data (hypothetical) presented in Table 13.1 illustrates the computation of *RL*. In computing both the *DL* and the *RL*, the stimulus sequence is varied with a minimum change in its magnitude in each presentation. Hence, Guilford (1954) prefers to call this method the *method of minimal changes.*

Table 13.1 Determination of the lower threshold of sound intensity by the method of limits

Sound Intensity (Arbitrary Scale Units)	A	D	A	D	A	D	A	D	A	D
15		+						+		+
14		+		+				+		+
13		+		+		+		+		+
12		+		+		+		+		+
11		+		+		+		+		+
10		+		+	+	+		+		+
9		+	+	+	–	+	+	+	+	+
8		+	–	–	–	+	–	+	–	+
7	+	+	–		–	–	–	–	–	–
6	–	–	–		–		–		–	
5	–		–		–		–		–	
4	–		–		–		–		–	
3	–		–		–		–		–	
2	–				–					
1	–									
Individual Thresholds:	6.5	6.5	8.5	8.5	9.5	7.5	8.5	7.5	8.5	7.5

Mean ascending threshold (MAT) $= \dfrac{6.5 + 8.5 + 9.5 + 8.5 + 8.5}{5} = 8.3$

Mean descending threshold (MDT) $= \dfrac{6.5 + 8.5 + 7.5 + 7.5 + 7.5}{5} = 7.5$

Mean absolute threshold $= \dfrac{\text{MAT} + \text{MDT}}{2} = \dfrac{8.3 + 7.5}{2} = 7.9$

SD $= 0.9165 = 0.92$

Table 13.1 illustrates the data taken in 10 series out of which 5 are in the ascending series and 5 are in the descending series. The first series is the ascending series and the experimenter starts the C_o (the sound) at 1 scale unit (which is arbitrary), and the subject reports 'No' (that is, he does not hear it) and accordingly, it is indicated by a minus sign in the series. Then, the experimenter goes on increasing the sound by one unit per trial and the subject continues to report 'No' until the experimenter reaches the seventh unit where the subject changes his response from 'No' to 'Yes' (that is, he hears it) and accordingly, it is indicated by a plus sign in the series. Thus the threshold for this particular series lies somewhere between 6 and 7 and hence, the midpoint, 6.5 becomes the threshold and is written below this column as one estimate of *RL*. Next comes the descending series where the experimenter starts the C_o stimulus at 15 units, which is much above the threshold found in the previous series. The subject reports 'Yes' which is indicated by a plus sign. Again, here the experimenter goes on decreasing the C_o by one unit in each trial until the subject changes his response. This time he changes his response from 'Yes' to 'No' at the sixth unit. The midpoint of 6 and 7 is 6.5 which becomes the threshold for this descending series and it is entered below this column as another estimate of *RL*. The thresholds thus found in each series are also called *transition points* above which the subject changes his response in the ascending series and below which he also changes his response in the descending series. Such transition points have been demarcated in Table 13.1. Thus, several alternate ascending and descending series are taken until the experimenter is well satisfied with the relative uniformity of the different individual thresholds. In each series the experimenter deliberately changes the starting point so that the subject does not fall into a habit of routine guessing, which may facilitate his task. For computing *RL* for the data obtained in 10 series, the individual thresholds entered at the bottom of each series are averaged; in our case it is 7.9. The standard deviation of this distribution is 0.92, which indicates variability in the subject's performance due to some variable errors like changes in his motivation, interest or attention, etc.

Besides these variable errors, the *RL* may also be affected by two constant errors—*the error of habituation* and *the error of anticipation*. The *error of habituation* (sometimes called the *error of perseverance*) may be defined as a tendency of the subject to go on saying 'Yes' in a descending series or 'No' in an ascending series. In other words, when the error of habituation is in operation, the subject falls into a habit of giving certain responses even after a clear change in the stimulus has occurred. One natural consequence of this error is to inflate the mean of the ascending series over the mean of the descending series. The *error of anticipation* (sometimes also called the *error of expectation*) is the opposite of the error of habituation and accordingly, may be defined as the tendency to expect a change from 'Yes' to 'No' in the descending series and 'No' to 'Yes' in the ascending series before the change in stimulus is apparent. The conscious tendency that works in the mind of the subject when such an error is in operation, is that he has said 'Yes' many times and, therefore, should now say 'No' (in the descending series); likewise, he thinks that he has said 'No' many times and, therefore, he should now say 'Yes' (in the ascending series). The natural consequence of such an error is to inflate the mean of the descending series over the mean of the ascending series. The primary purpose of giving alternate ascending and descending series is to cancel out these two types of constant errors. Table 13.1 shows that the mean of the ascending series is 8.3, which is higher than the mean of the descending series (that is, 7.5). It means that the subject has committed the error of habituation though its magnitude is very small, that is, 0.8 only. Since these two types of constant errors work in opposite directions, both cannot exist within the same subject. Practice and fatigue may affect the data obtained by the method of limits. These two effects may be shown by comparing the mean of the first half with the mean of the second half of the total series taken. Guilford (1954) has recommended that these effects can be analyzed in an even better way by the ANOVA (Analysis of variance).

Table 13.2 Determination of the difference threshold for weightlifting by the method of limits (data are hypothetical)

Comparative Stimuli (weight in g)	A	D	A	D	A	D	A	D	A	D
110		+				+				
109		+		+		+		+		+
108	+	+		+		+		+	+	+
107	=	+	+	+	+	+	+	+	=	+
106	=	=	=	+	=	=	=	+	=	+
$S_t = 105$	–	=	=	=	=	=	=	=	–	=
104	–	–	–	=	–	–	–	=	–	–
103	–		–	–	–		–	–	–	
102	–		–		–		–		–	
101	–		–				–			
100	–									
Upper Threshold:	107.5	106.5	106.5	105.5	106.5	106.5	106.5	105.5	107.5	105.5
Lower Threshold:	105.5	104.5	104.5	103.5	104.5	104.5	104.5	103.5	105.5	104.5
Mean:	106.5	105.5	105.5	104.5	105.5	105.5	105.5	104.5	106.5	105.5

$$\text{Mean upper threshold} = \frac{1064}{10} = 106.4$$

$$\text{Mean lower threshold} = \frac{1045}{10} = 104.5$$

$$\text{Interval of uncertainty} = 106.4 - 104.5 = 1.9$$

$$\text{Difference limen} = \frac{\text{Interval of uncertainty}}{2} = \frac{1.9}{2} = 0.95$$

$$\text{Point of subjective equality} = \frac{106.4 + 104.5}{2} = 105.45$$

$$\text{Constant error} = PSE - S_t = 105.45 - 105 = 0.45$$

Mean ascending series = 105.9

Mean descending series = 105.0

Table 13.2 illustrates the calculation of *DL* of the data obtained in an experiment on weightlifting (the data are hypothetical). The S_t is set at 105 g and there are 10 C_o stimuli, 5 below the S_t and 5 above the S_t. All C_o stimuli are of the same size and of the same texture. There are 10 series in which 5 are ascending and 5 are descending series. The first trial is of the ascending series where the experimenter starts with a C_o of 100 g weight, which is much below the S_t. The subject reports it to be lighter than the S_t and hence, the experimenter records it by putting a sign of minus. In this way, the C_o is increased by one gram weight in each presentation.

The first change in response by the subject from 'lighter' to 'equal' occurs at 106 g weight and accordingly, this is noted down by putting the equality sign. The experimenter does not stop here but continues to increase the weight of C_o by one gram until the subject changes his response for the second time. After two presentations, he reports the change in response from 'equal' to 'heavier' at 108 g weight which is noted by the experimenter by putting down a plus sign. The first change in response from the 'lighter' to 'equal' is the lower threshold of change and the second change in response from 'equal' to 'heavier' is the upper threshold of change. Like *RL*, the threshold of change is the midpoint of the two stimuli between which value change occurs. Thus the upper threshold is 107.5 (the midpoint of 107 and 108) and the lower threshold is 105.5 (the midpoint of 105 and 106). The same procedure is repeated in descending series except that the experimenter starts presenting the C_o from 110 g, which is clearly judged as 'heavier' than the S_t and accordingly, noted by putting down a plus sign. The first change in response from 'heavier' to 'equal' in descending series (the second series) occurs at 106 g weight and the second change in response from 'equal' to 'lighter' occurs at 104 g weight. Accordingly, the upper threshold is 106.5 and the lower threshold is 104.5. The same procedure is repeated in other ascending and descending series. Table 13.2 shows that the entire range of comparison values may be divided into three parts: *an upper part* where the plus judgements dominate, *a lower part* where the minus judgements dominate, and *a middle part* where equal judgements dominate. The middle part where equal judgements dominate is the interval of uncertainty (*IU*) and covers two *DLs* or *JNDs* from minus to equal and from equal to plus in the ascending series, and from plus to equal and from equal to minus in the descending series. The *DL* is half the *IU*, which is 0.95 and *IU* is the difference between the mean of the upper threshold (106.4) and the mean of the lower threshold (104.5). A *DL* of 0.95 indicates that the subject can correctly perceive the difference between the C_o and the S_t if the former is increased or decreased by a margin of 0.95 g. The midpoint of *IU* (106.4 + 104.5/2 = 105.45) becomes the point of subjective equality (*PSE*). The difference between the PSE and the S_t indicates the constant error (*CE*), which may be either positive or negative. If the PSE is above the S_t, the *CE* is positive and if the PSE is below the S_t, the *CE* is negative. In Table 13.2 a positive *CE* of 0.45 has been obtained, which indicates that the subject on the average has *overestimated* the standard weight of 105 g to the tune of 105.45 g. A negative *CE* indicates *underestimation* of the standard stimulus and reversely, the *overestimation* of variable stimulus. The mean of ascending series, which is calculated on the basis of the mean of the upper and the lower threshold of each ascending series, is higher than the mean of the descending series by 0.9 g. This difference indicates that the judgements of the subject have been influenced by the error of habituation to that extent. Since the difference is negligible (that is 0.9 only), it can be assumed that the judgements are not being influenced in a significant manner.

Method of Constant Stimuli (also known as Method of Right and Wrong Cases or Method of Frequency)

In this method, a number of fixed or constant stimuli are presented to the subject several times in a random order. The method of constant stimuli can also be employed for determining the *RL* or *DL*. For determining *RL*, the different values of the stimulus are presented to the subject in a random order (and not in an order of regular increase or decrease as is done in the method of limits), and he has to report each time whether he perceives or does not perceive the stimulus. Though the different values of stimulus are presented irregularly, the same values are presented throughout the experiment a large number of times, usually from 50 to 200 times each, in a predetermined order unknown to the subjects. The mean of the reported values of the stimulus becomes the index of *RL*. The procedure involved is known *as the method of constant stimuli*. Table 13.3 illustrates the computation of *RL* by this method. For calculating *DL*, in each presentation the two stimuli (one standard and one variable) are presented to the subject simultaneously or in succession (Guilford 1954, 118). On each trial, the subject is required to say whether one stimulus is 'greater' or 'less' than the other. In case of uncertainty, he reports, 'Doubtful' or 'Equal' and in such a situation sometimes the experimenter forces him to guess in order to avoid doubtful judgements. The procedure involved is known as the method of constant stimulus differences and not the method of constant stimuli. If the S_t and the C_o are to be presented in succession for half of the trials, the S_t is presented first and for the remaining half, the order is reversed. This is done to control a constant error (that is, time error), which may occur if the S_t is presented either before or after the C_o throughout the trials. In this sense, the method of constant stimulus differences differs from the method of limits where the C_o are presented by regularly increasing (ascending series) and decreasing (descending series) their value. In the method of constant stimuli, a smaller value of the C_o may be abruptly followed by a larger value of S_t or vice versa. The subject cannot estimate the likely C_o to be given for making a judgment. Though the different C_o are presented in an irregular order, they remain constant throughout the presentations in the experiment, that is, the same value is presented throughout all presentations in the experiment in an irregular order. This method is also known as *the method of right or wrong cases* because in each case, the subject has to report whether he perceives the stimulus (right) or he does not perceive the stimulus (wrong). But this name is almost obsolete now. This method has one distinct advantage over the method of limits. Since in the method of limits the stimuli are presented in a regular increasing or decreasing manner, the two constant errors, that is, the error of habituation and the error of expectation, are inevitable. But these two errors are safely avoided in the constant methods because the presentations of the stimuli are random or irregular.

Table 13.3 illustrates the determination of *RL* from the hypothetical data taken from a hypothetical subject.

Table 13.3 Determination of the lower absolute threshold for sound intensity

Stimulus value (Herztone)	Frequencies showing the number of times tone heard	Percentage of times tone heard
50	98	98
48	80	80
46	70	70
44	65	65
42	43	43
40	30	30
38	20	20
36	15	15
34	12	12
32	3	3

In this hypothetical experiment which yielded the data of the Table 13.3, the subject was presented 10 sound intensities, each 100 times. He was to report whether or not he heard the sound. Doubtful judgements were not allowed. Against each sound intensity are given the frequencies with which the subject reported the sound being heard clearly. (Table 13.3 shows only the frequencies with which each sound intensity was perceived). Subsequently, the frequencies were converted into the percentage. Since there were 100 judgements for each stimulus intensity, the number of times the sound intensity was heard and the percentage of the time the sound was heard were identical. There are several methods for calculating *RL*, out of which the three most common ones will be discussed here. These three methods are the graphical method, *the linear interpolation method* and the *summation method*. The graphical method is a rough method of calculating *RL*. Here the stimulus dimensions are located on the abscissa and the percentages showing the number of times the stimulus was perceived are located on the ordinate of the graph. As we know, *RL* is that point of the stimulus dimension which is perceived 50% of times and not perceived in the remaining 50% of times. Accordingly, we go from the 50% point on the ordinate across to the curve and then a straight line is dropped on the abscissa. The *RL* is assumed to be located on the point where the perpendicular line hits the abscissa (See Figure 13.1) The *RL*, thus found, is 42.60 which is the point that produces response (report of hearing the sound) 50% of the time.

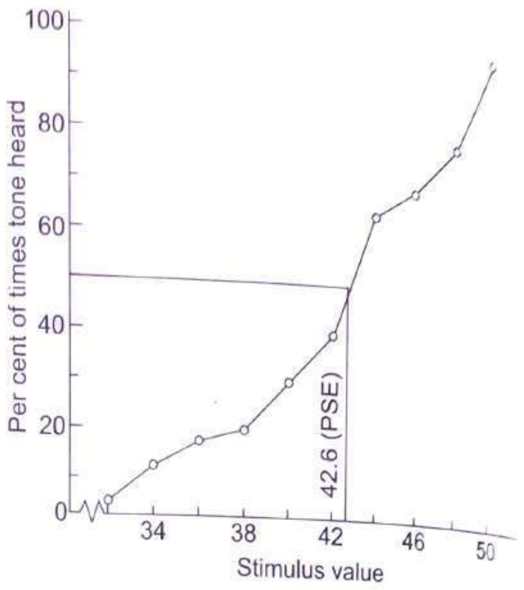

Fig. 13.1 Graphical determination of the absolu threshold for sound intensity

The linear interpolation method is more precise and exact than the graphical method. In this method, the median (or Q_2) of the distributions of judgment is found by interpolation. While calculating *RL* from the data given in Table 13.3, our primary interest is to know that stimulus value which cuts the distributions of judgements into two equal halves so that above that point we get half of the increasing frequencies of the sound being heard, and below which we get the decreasing frequencies of the sound being heard. An inspection of Table 13.3 reveals that this point falls in between the stimulus value of 42 (43%) and 44 (65%). The difference between 65 and 43 is 22 and if we go 7 points above the stimulus values of 43, we get 50% because 43 plus 7 is 50. By interpolation, then, we calculate $(7/22) \times 2$, because there are two step intervals between the stimulus value of 42 and 44. The resulting value is 0.636 which, if added to 42, becomes 42.636 or 42.64. Thus the *RL*, by the linear interpolation method is 42.64, which is approximately equal to 42.60 found by the graphical method. The method assumes a straight line (or linear) relationship between the two stimulus values used in the interpolation and hence, it is called the *method of linear interpolation*.

The linear interpolation method has been criticized on some grounds. *First*, the assumption of the method that there exists a linear relationship between the two stimulus values has been a controversial one. *Second*, the method does not fully utilize all data in computation of the median point. In the above example, only two of ten judgements, namely, 43% and 65% have been used. *Third*, no correct estimate of dispersion of judgements can be made in this method. As a consequence, the dependability upon *RL* thus calculated is poor. Guilford (1954) has, however, proposed an indirect way of calculating the standard deviation (a measure of dispersion) through *Q* or the semi-interquartile range from data obtained by the method of linear interpolation. The step consists in finding the stimulus value corresponding to $p = 0.25$ (or Q_1) and $p = 0.75(Q_3)$. The one half of difference between Q_3 and Q_1 becomes *Q* or the semi-interquartile range and assuming the distribution to be a normal one, $\text{SD} = 1.4833 \times Q$.

The summation method proposed by Woodworth (1938) is another method of calculating *RL*. This method is most suited where the terminal proportions can be expressed as 0 and 1.00 or when it can be assumed that the next terminal proportions are close to 0 and 1.00. In the method of summation, the mean (which shows the *RL*) and the *SD* (which shows dispersions of judgements) can be estimated by the equations given below:

$$M = S_1 - 0.5i - i \text{ sum } p \qquad (13.2)$$

$$M = S_0 + 0.5i + i \text{ sum } d \qquad (13.3)$$

where, S_1 = stimulus value where p = 1.00 (in case p is not 1.00, it can be assumed at the next stimulus value to be so);

i = step interval of the stimulus value;

p = proportions of judgement in higher or superior judgement category;

$d = 1 - p$ (or proportions of judgements in lower or inferior judgment category);

S_0 = stimulus value where $d = 0$ (in case, d is not 0, it can be assumed at the next stimulus value to be so).

A close inspection of equations 13.2 and 13.3 reveals that Equation 13.2 estimates the mean on the basis of p and Equation 13.3 estimates the mean on the basis of d. Since p and d are simply two ways of expressing the same proportion, the mean calculated by the two methods should be exactly the same. Some investigators have preferred to write Equation 13.2 as Mean = apical value $-i/2 - i$ sum p and Equation 13.3 as Mean = basal value $+ i/2 + i$ sum d, where apical value is defined as that stimulus value where all responses given by the subject belong to the same category and below which the responses are divided into categories such as 'Heavier' or 'Lighter', 'Heard', or 'Not heard', 'Yes' or 'No', etc., and the basal value is that stimulus value where all responses given by the subject belong to the same category and above which they are divided into different categories. The meaning of the rest of the symbols is as usual. When mean is calculated by Equation 13.2, the *SD* should be calculated by Equation 13.4 where

$$\text{SD} = i\sqrt{2 \text{ sum } cp - (\text{sum } p)^2 - \text{sum } p} \qquad (13.4)$$

When mean is calculated by Equation 13.3, the *SD* should be calculated by Equation 13.5:

$$\text{SD} = i\sqrt{2 \text{ sum } cd - (\text{sum } d)^2 - \text{sum } d} \qquad (13.5)$$

In Equation 13.4 the meaning of the symbols are defined as in Equation 13.2 except cp which has proportions p cumulated from low to high values of p, and similarly, in Equation 13.5 the symbols are defined as usual except cd which has proportions d cumulated from low to high values of d. Table 13.4 illustrates the calculation of *RL* by the summation method.

Table 13.4 Determination of lower absolute threshold for sound intensity by the summation method (data are replicated from Table 13.3)

Stimulus value (Herztone)	p (proportion of time heard)	d (proportion of time not heard)	cp	cd
50	0.98	0.02	4.36	0.02
48	0.80	0.20	3.38	0.22
46	0.70	0.30	2.58	0.52
44	0.65	0.35	1.88	0.87
42	0.43	0.57	1.23	1.44
40	0.30	0.70	0.80	2.14
38	0.20	0.80	0.50	2.94
36	0.15	0.85	0.30	3.79
34	0.12	0.88	0.15	4.67
32	0.03	0.97	0.03	5.64
Sum	4.36	5.64	15.21	22.25

$S_1 = 52$ (assumed), $S_0 = 30$(assumed)

$$M = S_1 - 0.5i - i \text{ sum } p$$
$$= 52 - (0.5)(2) - (2)(4.36)$$
$$= 52 - 1 - 8.72$$
$$= 42.28$$

Similarly,

$$M = \text{apical value} - i/2 - i \text{ sum } p$$
$$= 52 - 2/2 - 2(4.36)$$
$$= 52 - 1 - 8.72$$
$$= 42.28$$

$$M = S_0 + 0.5i - i \text{ sum } d$$
$$= 30 + (0.5)(2) + (2)(5.64)$$
$$= 30 + 1 + 11.28$$
$$= 42.28$$

Similarly,

$$M = \text{basal value} + i/2 + \text{sum } d$$
$$= 30 + 2/2 + (2)(5.64)$$
$$= 30 + 1 + 11.28$$
$$= 42.28$$

$$SD = i\sqrt{2 \text{ sum } cp - (\text{sum } p)^2 - \text{sum } p} = 2\sqrt{2(15.21) - (4.36)^2 - 4.36} = 5.31$$

$$SD = i\sqrt{2 \text{ sum } cd - (\text{sum } d)^2 - \text{sum } d} = 2\sqrt{2(22.25) - (5.64)^2 - 5.64} = 5.31$$

It is obvious that the *RL* calculated by the summation method is very close (except the fractional differences which may be due to the rounding errors) to *RL* calculated by the graphical method (42.60) and the linear interpolation method (42.64), and this checks the accuracy of the calculation.

The calculation of *DL* may also be done through the graphical method, the linear interpolation method and the summation method. Suppose the experimenter wants to determine the *DL* for lifted weights in which three categories of judgements (heavier, lighter and equal) are obtained. Let the S_t be 100 g and the C_o be 70 g, 80 g, 90 g, 100 g, 110 g, 120 g and 130 g. Each of these C_os is paired with S_t 100 times in a random order and the subject is instructed to judge whether the second stimulus is heavier, lighter or equal to the first one. The resulting frequencies and proportions of the judgements have been shown in Table 13.5. (The experiment is hypothetical.)

Table 13.5 Frequencies and proportions of three categories of judgements in a weightlifting experiment where standard stimulus is 100 g

Stimulus values	Frequencies			Proportions		
	Heavier	Equal	Lighter	*p* Heavier	*m* Equal	*d* Lighter
130	90	8	2	0.90	0.08	0.02
120	84	12	4	0.84	0.12	0.04
110	78	18	4	0.78	0.18	0.04
S 100	40	50	10	0.40	0.50	0.10
90	35	20	45	0.35	0.20	0.45
80	10	10	80	0.10	0.10	0.80
70	0	1	99	0	0.01	0.99
$S_1 = 140$ (assumed)		$S_0 = 60$ (assumed)		$\Sigma = 3.37$	1.19	2.44

The distribution of the judgements—'heavier', 'equal' and 'lighter'—has been shown in Figure 13.2, which may be considered as a typical distribution in the field of sensitivity. For calculating *DL* by the graphical method, two curves (technically called the psychometric function) have been drawn in Figure 13.3.

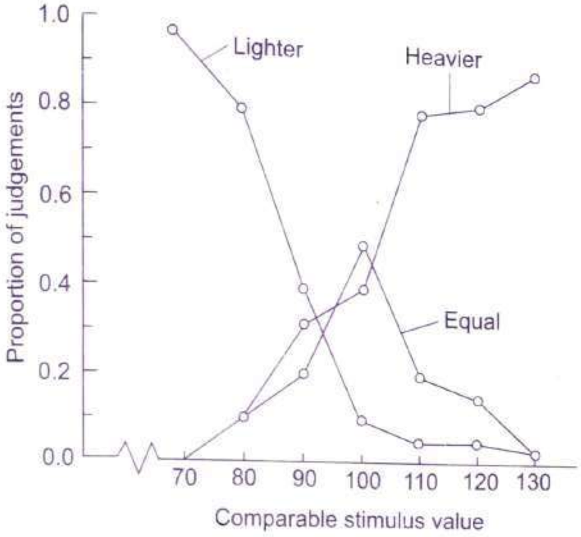

Fig. 13.2 Distribution of three categories of judgement made by subjects

One psychometric function or curve is for the judgements of 'heavier', and another psychometric function or curve is for the judgements of 'lighter'.

These two curves are shown as ascending functions of the proportions of the two types of the judgements and divide the areas representing the three categories of judgements into three separate regions (see Figure 13.3). It is obvious from Figure 13.3 that the upper threshold or limen (L_u) is 102.60 and the lower threshold or limen (L_l) is 88.60.

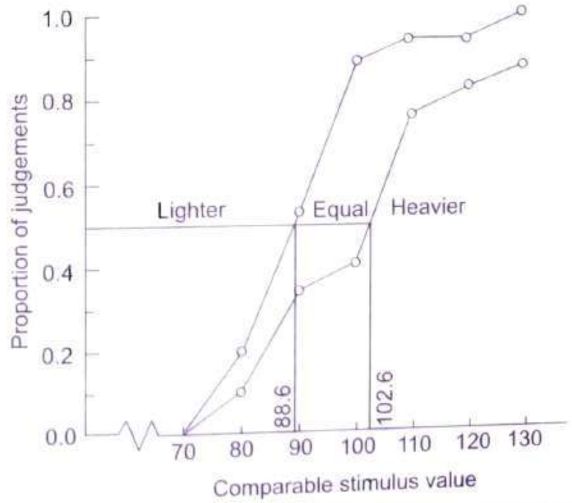

Fig. 13.3 Areas representing the three categories of judgement which divide the transition zones into three categories

These two limits mark the interval of uncertainty or *IU* and therefore, *DL* is $(102.60 - 88.60)/2 = 7$ and the point of subjective equality of *PSE* is $(102.6 + 88.6)/2 = 95.6$. Similarly, by the linear interpolation method the upper limen which is expected to fall between the stimulus value 100 and 110 (because the proportion 0.50 lies therein), can be calculated like this: The difference in the proportions of 'heavier responses' of these two limits of stimulus value is 0.38 and if 0.10 is added to 0.40, we get 0.50. Hence, $(0.10/0.38) \times 10 = 2.63$ which is added to 100 to yield 102.63, the upper limen (L_u) we require. Likewise, the lower limen (L_l), that is, of 'lighter responses', is expected to fall in between the stimulus value 90 and 80. By interpolation, we get $(0.05/035) \times 10 = 1.42$ which if subtracted from 90 leaves 88.58, the value of the lower limen we require. Then *DL* is $= (102.63 - 88.58)/2 = 7.025$ and *PSE* is $= (102.63 + 88.58)/2 = 95.605$. These values are approximately equal to those found by the graphical method and hence, check the accuracy of the calculation. The *DL* can also be calculated by Woodworth's summation method. The upper limen by Equation 13.2 is $140 - (0.5)(10) - (10)(3.37) = 101.3$ and the lower limen, by Equation 13.3, is equal to $60 + (0.5)(10) + (10)(2.44) = 89.4$. The *DL* is $(101.3 - 89.4)/2 = 5.9$ or 6 and the PSE is $(101.3 + 89.4)/2 = 90.34$. The *DL* can also be estimated from the sum of the proportions of 'equal' or 'doubtful' judgements. Here, $IU = i$ sum m and *DL*, as usual, is half of *IU*. In Table 13.5, *IU* is $10 \times 1.19 = 11.9$, the *DL* being half of this which becomes 5.95 or 6. It is obvious that some minor variations in the value of *DL* and *PSE* occur when different methods of calculation are adopted.

The *DL* can also be estimated from the two-category judgement. Suppose the same weightlifting experiment was carried on a new subject who gave his judgement with respect to the second stimulus as being 'heavier' or 'lighter' only. Sometimes, it happens that even when the subject is forced to give a two-category judgement, he reports 'doubtful' or 'equal' responses. In such a situation, the ideal approach is to divide the number of 'doubtful' or 'equal' responses equally between the two categories such as 'heavier' or 'lighter' at each stimulus value. The rationale behind this is that the subject gives 'doubtful' or 'equal' responses only in situations which are equally inclined towards 'heavier' or 'lighter'. Others however, reject this view and prefer to distribute the doubtful judgements in proportion to the numbers of judgements already made towards the two categories. This is known as the proportional division of 'equal' or 'doubtful' judgements. In this experiment, let us suppose that no 'doubtful' or 'equal' responses were made by the subject and hence, the distribution of the frequencies and proportions resulting from 100 pair presentations of each C_o with S_t was as given in Table 13.6.

Table 13.6 Distribution of the frequencies and proportions of two-category judgements in a weightlifting experiment where the standard stimulus is 100 g

Stimulus values	Frequencies		Percentage	
	Heavier	Lighter	Heavier	Lighter
130	98	2	98	2
120	85	15	85	15
110	70	30	70	30
S 100	62	38	62	38
90	40	60	40	60
80	10	90	10	90
70	1	99	1	99

The reason why experimenters, in general, prefer two-category judgements to three-category judgements is due to certain advantages of the former over the latter (Postman & Egan 1949). The first advantage is that in the two-category judgement, where the subject is forced to guess by restricting his judgement to either 'greater' or 'less', he will be more often right than wrong in his guessing.

The second advantage is that when a neutral category like 'equal' is used, the meaning of this category becomes vague and confusing. Not only this, but its meaning also varies from subject to subject. When the subject is forced to give his judgement in either of the two categories, such confusion and vagueness are automatically removed. In view of these advantages, it is suggested in general, that the two-category judgement be used.

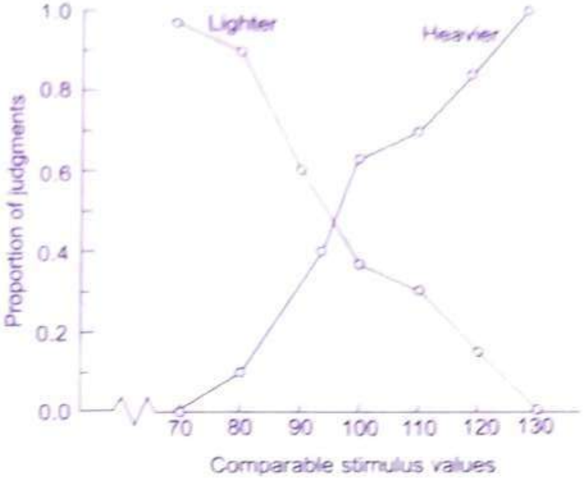

Fig. 13.4 Distribution of two-category judgement made by subjects

In computing *DL* from a two-category judgement by the graphical method and the linear interpolation method, the statistical treatments are different from those used in the three-category judgement. Here the linear interpolation method involves the calculation of Q_1 and Q_3. Q_1 and Q_3 will be the lower difference threshold and the upper difference threshold respectively. The range between Q_1 and Q_3 is called the interquartile range and is equivalent to *IU*. As *DL* is half of *IU*, the range between Q_1 and Q_3, if divided by 2, yields the value of *Q* (or the semi-interquartile range) or the difference threshold. Q_1 (or 25th percentile) from the data of Table 13.6 may be calculated like this:

$$Q_1 = 80 + 10\frac{25-10}{40-10} = 80 + 10\frac{15}{30} = 80 + 10(0.5) = 85$$

$$Q_3 = 110 + 10\frac{75-70}{85-70} = 110 + 10\frac{5}{15} = 110 + 10(0.333) = 113.33$$

Here, 110 is the stimulus value of the step above which Q_3 (or the 75th per cent) falls; 10 is the step interval; 75 is the per cent required for Q_3; 85 is the per cent for that stimulus value where Q_3 lies; and 70 is the per cent for that stimulus value above which Q_3 lies. A similar interpretation is applicable for those values on the basis of which Q_1 is calculated. Now *DL* and *PSE* can be calculated as usual:

$$DL = \frac{\text{upper threshold} - \text{lower threshold}}{2} \text{ or } \frac{\text{IU}}{2}$$

$$= \frac{113.33 - 85}{2} = \frac{28.33}{2} = 14.165$$

$$PSE = \frac{\text{upper threshold} + \text{lower threshold}}{2} = \frac{113.33 + 85}{2} = 99.165$$

The *DL*, by the graphical method, is calculated by drawing a freehand smoothed curve which passes close to the data points, leaving some of them above and some of them below the curve and dropping two perpendiculars at the abscissa from 25% (or 0.25) and 75% (0.75) points located on the *y*-axis and joined by two separate horizontal lines (see Figure 13.5). The distribution of the judgements 'heavier' and 'lighter' has been shown in Figure 13.4.

It is obvious from Figure 13.5 that the perpendicular dropped from 75% touches the abscissa at 113.4 g, which becomes the upper difference threshold (L_u) and the perpendicular dropped from 25% touches the abscissa at 84.6g, which becomes the lower difference threshold (L_l). As usual, the $DL = (L_u - L_l)/2 = (113.4 - 84.6)/2 = 14.4$ and the $PSE = (113.4 + 84.6)/2 = 99$. The values of *DL* and *PSE* arrived at by the graphical method are very close to those obtained by the linear interpolation method. Sometimes the perpendicular dropped from 50% falls exactly on the value of the standard stimulus at the abscissa (which is not the case in Fig. 13.5). In such a situation, the constant error is zero. Thus the perpendicular dropped from 50% at abscissa indicates the value of *PSE*. If the constant error is positive, the perpendicular dropped from 50% will lie to the right of the value of the standard stimulus shown at abscissa and if the constant error is negative, the perpendicular dropped from 50% will lie to the left of the value of the standard stimulus at the abscissa. In this example, the constant error is equal to −1. Hence, the perpendicular drawn from 50% lies to the left of the value of standard stimulus shown at abscissa.

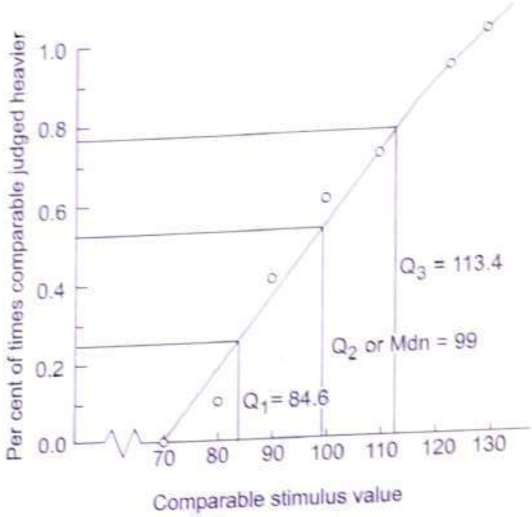

Fig. 13.5 Calculation of *DL* and *PSE* by the method of constant stimuli

Method of Average Error (also known as Method of Adjustment, Method of Reproduction or Method of Equivalent Stimuli)

The method of average error is the oldest method of psychophysics and is a sort of gift to psychophysics and astronomy. In this method, the subject is provided with an S_t and a C_o. The C_o is either greater or lesser in intensity than the S_t. He is required to adjust the C_o until it appears to him to be equivalent to the S_t.

The difference between S_t and C_o defines the error in each judgement. A large number of such judgements are obtained and the arithmetic mean (or average) of those judgements is calculated. Hence, the name, 'method of average error or mean error' is given. The obtained mean is the value for *PSE*. The difference between the S_t and *PSE* indicates the presence of the constant error, or *CE*. Equation 13.1 may thus be applied for calculating the *CE*. If the *PSE* (or average adjustment) is larger than the S_t, *CE* is positive and indicates *overestimation* of the standard stimulus. On the other hand, if the *PSE* is smaller than the S_t, *CE* is negative and indicates *underestimation* of the standard stimulus.

The method of average error is distinguished from the two preceding methods in one important way. In the preceding two methods, the control over changes in the stimulus was entirely in the hands of the investigator or the experimenter. But in the method of average error, the subjects themselves are permitted to control the variations or changes in the stimulus. In using the method of average error, care must be taken to see that the probability of systematic or constant error as well as variable error is minimised. This can be ensured by the following means:

1. In half of the total trials the C_o should be set at a value larger than the S_t and in the remaining half, the C_o should be set at a value smaller than the S_t. In this way, the direction of adjustment should be counterbalanced so that movement error may be minimized or cancelled. The *movement error* is the error produced by the subject's bias for moving the comparable stimulus inward or outward.

2. The spatial presentation of the C_o and the S_t may also result in a systematic error called 'space error'. The *space error* is defined as the error which is produced by the subject's bias in adjusting the C_o with S_t when the former is placed either to the left or right of the latter in all the trials. Hence, for controlling space error, the C_o should be presented by placing it to the right of the S_t in half of the total trials and reversing its position in the remaining half.

3. The initial value of the C_o should be randomly changed from trial to trial so that the subject may not get any unnecessary cues known as 'extraneous cues' in adjusting or equating the C_o to the S_t.

The method of average error is also known by several other names. In this method, the subjects adjust the C_o to the S_t by making an active manipulation of the C_o. Hence, the method is also known as the 'method of adjustment'. The purpose of the method is to determine equivalent stimuli by active adjustment of the C_o by the subject in each trial and hence, the method also goes by the name of 'method of equivalent stimuli'. In this method, the subject tries to reproduce a given C_o in a way which may seem equivalent to the S_t. Hence, it is also known as the 'method of reproduction'.

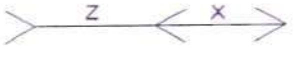

Fig. 13.6 Muller-Lyer illusion

The main purpose of this method is to calculate *PSE*, although *DL* can also be calculated. The method will be illustrated with data obtained from the experiment on the Muller-Lyer illusion (data are hypothetical). The Muller-Lyer illusion has been propounded by a German sociologist-cum-psychiatrist Franz Carl Muller-Lyer (1857–1916). A figure producing the Muller-Lyer illusion is given in Figure 13.6.

In Figure 13.6, line Z is known as the feather-headed line and line X is known as the arrow-headed line. Commonly, the arrow-headed line is the fixed or standard stimulus and the feather-headed line is the comparable or variable stimulus. The experimenter sets the variable stimulus, that is, the feather-headed line in a way that it appears clearly shorter or longer than the arrow-headed line. The task of the subject is to adjust the feather-headed line until it appears phenomenally equal to the arrow-headed line. When the experimenter sets the feather-headed line shorter than the standard arrow-headed line, the subject must move it outward in order to make it equal to the standard. This is called the *outward movement*. When the experimenter sets the feather-headed line longer than the arrow-headed line, the subject must move it inward. This is called *inward movement*.

In an experiment on the Muller-Lyer illusion, there are two common constant errors committed by the subject. They are the *movement error* and *space error*. Movement error occurs when the subject has a certain bias for one of the two movements, that is, inward movement and outward movement, which unduly helps him in making the feather-headed line equal to the arrow-headed line. Similarly, space error results when the subject has a certain bias for one of the two spaces in the visual field, namely, right and left, which helps in adjusting the feather-headed line. Some subjects may get facilitation in adjusting the feather-headed line when it is placed in the right visual field and some subjects may get the same type of facilitation when the feather-headed line is placed in the left visual field. Both these facilitations tend to produce a space error. The movement error and the space error should be controlled in the experiment. One convenient method of controlling the two types of constant errors is to use the counterbalancing design in which an equal number of settings or trials will be given for both the types of movements (inward movement and outward movement) and for both the types of spaces (right visual field and left visual field). Let O stand for outward movement, I stand for inward movement, L stand for left visual field (where the feather-headed line is placed left of the arrow-headed line) and R stand for right visual field (where the feather-headed line is placed right of the arrow-headed line). Following the counterbalancing design, a set of the 64 trials may be distributed as follows:

Order:	A	B	B	A
	L OIIO (16 trials)	R OIIO (16 trials)	R OIIO (16 trials)	L OIIO (16 trials)

The raw data (hypothetical) are presented in Table 13.7.

Table 13.7 Judgements (in MM) of the length of the feather-headed line made by a single subject in 64 trials—calculation of the extent of illusion shown

Trial	Direction	Length judged	Trial	Direction	Length judged	Trial	Direction	Length judged	Trial	Direction	Length judged
1	LO	48	17	RO	45	33	RO	51	49	LO	48
2	LI	47	18	RI	43	34	RI	50	50	LI	49
3	LI	45	19	RI	46	35	RI	50	51	LI	47
4	LO	47	20	RO	45	36	RO	48	52	LO	46
5	LO	46	21	RO	46	37	RO	49	53	LO	46
6	LI	45	22	RI	48	38	RI	47	54	LI	47
7	LI	44	23	RI	43	39	RI	46	55	LI	48
8	LO	51	24	RO	44	40	RO	46	56	LO	49
9	LO	53	25	RO	45	41	RO	45	57	LO	50
10	LI	49	26	RI	48	42	RI	43	58	LI	51
11	LI	48	27	RI	47	43	RI	44	59	LI	52
12	LO	47	28	RO	47	44	RO	47	60	LO	50
13	LO	46	29	RO	46	45	RO	46	61	LO	49
14	LI	47	30	RI	48	46	RI	43	62	LI	48
15	LI	45	31	RI	49	47	RI	45	63	LI	45
16	LO	44	32	RO	50	48	RO	46	64	LO	44

Mean of *LO* Series = 47.75 Mean of *LI* Series = 47.31

Mean of *RO* Series = 46.62 Mean of *RI* Series = 46.25

$$\text{Space error} = \frac{(\textit{Mean of LO} + \textit{Mean of LI})}{2} - \frac{(\textit{Mean of RO} + \textit{Mean of RI})}{2}$$

$$= \frac{(47.75 + 47.31)}{2} - \frac{(46.62 + 46.25)}{2} = 47.530 - 46.435 = 1.095$$

$$\text{Movement error} = \frac{(\textit{Mean of LO} + \textit{Mean of RO})}{2} - \frac{(\textit{Mean of LI} + \textit{Mean of RI})}{2}$$

$$= \frac{(47.75 + 46.62)}{2} - \frac{(47.31 + 46.25)}{2} = 47.185 - 46.780 = 0.405$$

Mean of 64 judgements or $PSE = 46.98$, SD of 64 judgements $= 2.315$

Constant error (or the extent of illusion): $PSE - \textit{Standard stimulus} = 46.98 - 50 = -3.02$

It is obvious from Table 13.7 that *PSE* is 46.98. Since the length of standard stimulus (the arrow-headed line) is 50 mm, the CE is −3.02. This value of CE indicates the underestimation of the standard arrow-headed line. In other words, the subject, on the average, perceives the arrow-headed line to be shorter than the feather-headed line to the extent of 3.02 mm. Thus the extent of the Muller-Lyer illusion in the hypothetical data of Table 13.7 is 3.02. The evidence for the space error and the movement error has also been found, although these values are not large. The space error is 1.095 and the movement error is 0.405. From the data it is obvious that the subject has some bias, which facilitates the adjustment of the feather-headed line when it is placed to the left rather than the right of the standard arrow-headed line. Similarly, the subject appears to be more biased towards outward movement than towards inward movement in adjusting the feather-headed line.

WEBER'S LAW AND FECHNER'S LAW

Weber's law, named after its discoverer E H Weber, was the first systematic attempt to formulate a principle which governed the relationship between psychological experience and physical stimulus. For a long time, this law has been the focus of psychophysical experimentation. The question that worried experimenters earlier was: Is the difference limen or *JND* (just noticeable discrimination) a fixed value for any given sense modality? This problem was thoroughly studied under the leadership of Weber who came to the conclusion that the *JND* was not a fixed value, rather it increased with the size of the standard stimulus in a linear fashion. In other words, as the magnitude of the standard stimulus is increased, the size of change needed for discrimination between the standard and the comparable stimulus (that is, *JND*) is also increased. Thus the greater the magnitude of the standard stimulus, the greater the size of the *JND* or *DL*. Weber's law is a common mathematical statement of this fact. This relationship between the size of the standard stimulus and the size of *JND* is technically known as Weber's Law. For example, if addition of one candle makes a just noticeable difference to an already lighted room having ten candles, it would take 10 candles to make the same difference in a lighted room having 100 candles and in a room lighted with 1000 candles, 100 candles will produce just the same noticeable difference. Thus, it is obvious that *JND* bears a constant ratio with the standard stimulus. In the words of Underwood (1966, 165), "The law states that for a given stimulus dimension, the *DL* bears a constant ratio to the point on the dimension (standard stimulus) at which the *DL* was measured." The law may be stated in terms of the following equation:

$$\frac{\Delta R}{R} = K \tag{13.6}$$

where, $\Delta R = DL$; $R =$ standard stimulus; and $K =$ constant.

Equation 13.6 may be verbally expressed as shown below:

$$\frac{DL}{\text{Standard stimulus}} = \text{constant} \tag{13.7}$$

The constant in Weber's law is always a fraction and is known as the Weber fraction or proportion. It indicates the proportion by which the standard stimulus must be increased in order to produce the just noticeable difference or to detect a change. If the addition of 8 g makes a difference in weight sensitivity of 10 g, the Weber fraction will be 8/10 = 0.8, which indicates that the intensity of the standard stimulus must be increased by 0.8 in order to perceive a difference between that stimulus and the other stimulus. If the Weber fraction is 0.8, a stimulus value of, say, 100 will be increased by 0.8 of its magnitude, that is by 80, that is, (0.8×100) to be just noticeably different from 100. In other words, it should be 180. Likewise, if the stimulus value is 1000, its value should be increased by 800 in order to produce the just noticeable difference. Thus the ratio remains constant irrespective of the strength of the standard stimulus.

Weber's law, in general, has been regarded as a good measure of the overall sensitivity in different sense modalities. Generally, if the Weber fraction is larger, the *DL*s for the given stimulus dimensions will also be larger. One advantage of the Weber fraction is its direct comparability. Since the fraction is not dependent upon a physical unit in terms of which the standard stimulus and the *DL* are measured, the fractions can be compared across the different stimulus dimensions. Weber's fraction has been commonly reported to be 0.020 for heaviness, 0.030 for line length, 0.079 for brightness, 0.023 for finger span, 0.014 for electric shock, and 0.084 for a salty taste.

One general difficulty with Weber's law is that its precision is lost where the standard stimulus reaches the extremes, that is, when the standard stimulus becomes either very weak or very strong, the precision is lost to a great extent. Not only that, the Weber fraction is also influenced by the way the stimuli are presented (Ono 1979). When the presentation is such that the standard stimulus is followed by the comparable stimulus, the fraction reaches its maximum precision. When the order is reversed or modified, its precision is adversely affected. For these reasons, several alternative laws have been derived for fitting the function of *DL* and the standard stimulus. The reader is referred to Guilford (1954) for a detailed discussion of these laws.

Fechner derived his law (called the Fechner Law) from Weber's Law. Fechner's law is an indirect method of scaling* judgement where *DL* is used as the unit of the equal-interval scale. Fechner was of the view that *DL* for each successive unit or psychological step can be determined by using a constant multiple. For example, suppose the Weber fraction is 0.25 (or 1/4) for a particular sensation. If one stimulus value is 20 units the other stimulus should be 1/4 of 20 or $(0.25 \times 20) = 5$ units more than 20, that is, it should be 25 for producing a just noticeable difference. Again the other stimulus value at the next psychological step to be just noticeably different from 25 units should be $0.25 \times 25 = 6.25$ more than 25 units, that is, it should be 31.25. Likewise, at the next psychological step the other stimulus value should be $31.25 + 7.81 = 39.06$ in order to produce the just noticeable difference from the stimulus value of 31.25. Thus the stimulus value required for each successive step or psychological step in order to produce a just noticeable difference should be $1\frac{1}{4}$ times the preceding one. In this way for each successive unit, large increments in stimulus value are needed to produce equal increments in psychological sensations. This increase in psychological sensation as a function of increment in stimulus value can also be easily described by a logarithmic relationship because it entails multiplication by a constant. For example, in the above example for each successive step, the *DL* can also be found by using a constant multiple, 5/4, such as for the first step starting with the stimulus value of 20, it would be $\frac{5}{4} \times 20 = 25$; for the second step it would be $\frac{5}{4} \times 25 = 31.25$; and for the third step, another stimulus for having *DL* would be $\frac{5}{4} \times 31.25 = 39.06$, and so on. It is obvious that increments in the stimulus value occur by a process of multiplication but increments in the resulting psychological sensation at each successive step occur by the process of addition. The former type of increment is known as 'geometrical progression' and the latter type of increment is known as 'arithmetical progression'. When one of the two variables increases in geometrical progression and the other in arithmetical progression, the relationship is termed as a 'logarithmic relationship'. Fechner's law states that the stimulus values and the resulting psychological sensation have a logarithmic relationship so that when the former increases in geometrical progression, the latter increases in arithmetical progression. In other words, the law states that the magnitude of sensation (or response) varies directly with the logarithm of the stimulus value. This law of Fechner has been paraphrased by one author, " The sensation plods along step by step while the stimulus leaps ahead by ratios." (Woodworth 1938). Fechner has given several formulas for showing this relationship but the most common is:

---

* On the basis of the method adopted for obtaining judgments there can be two types of scaling—direct scaling and indirect scaling. Direct scaling is one where direct method for obtaining judgment regarding psychological quantities on the interval or the ratio scale or the ordinal scales are adopted. In such methods, the experimenter clearly states the quantitative properties desired by the subject in the instruction. He may, for example, give instruction to judge by how many grams the C_o appears heavier or lighter than the S_t. Indirect scaling is one where no such direct methods of obtaining judgments are available and, usually, the equal intervals of judgments are derived from the same proportions or ratios. Hence, indirect scales can operationally be called *Confusion scales (Kling and Riggs 1984).*

$$R = K \log S \tag{13.8}$$

where, R = magnitude of sensation of response; K = Weber's constant and S = the magnitude of stimulus value.

In formulating his above law, Fechner made two important assumptions.

1. The *DL* or *JND* indicates equal increments in psychological sensation irrespective of the absolute level at which it is produced.
2. Psychological sensation is the sum of all those *JND* steps, which come before its origin.

Fechner's law had been very influential, particularly in psychology's early days, because the law showed that it was possible to relate the things of the mind to those of the body in precise, quantitative terms. But later investigators raised some objections and posed the question: Is it really true that sensory magnitude can only be measured indirectly as Fechner had claimed?

STEVENS' POWER LAW

One of the important Harvard psychologists, S S Stevens put forward the argument that sensory magnitude can be assessed by a straightforward method and for it, indirect method, as devised by Fechner, is not the only method. Stevens adopted a method called *Method of Magnitude Estimation* in which sensory magnitude could be directly assessed. He presented his subjects with a series of stimuli to which they had to assign numbers which were proportional to the corresponding subjective impressions of the subject. For example, if a tone appears three times broader than another, the subject was required to assign a number to the first that was three times larger than the number given to the second. Suppose, for the second tone he has assigned 4, then he would assign 12 for the first tone.

There were two major findings of Stevens' work as given below.

(i) The subjects had little or no trouble in performing the task. They can readily judge their own subjective experiences, thereby, producing a direct scale of their subjective magnitude.

(ii) This second important finding was the particular relationship between this direct scale and physical intensity. According to Stevens, the function was not logarithmic as Fechner had proposed, rather it was a power function. In terms of equation, it was

$$S = KI^N \tag{13.8a}$$

where, S = subjective magnitude;

I = Stimulus intensity;

K and N are constants.

This exponential function is called the *Power Law* or after its popularizer, *Stevens' Power Law*. According to this law, the actual rate of sensation change with changes in stimulus intensity depends upon the size of the exponent, that is, *N*. In other words, the intensity of a sensation is proportional to stimulus intensity raised to a certain power.

According to Stevens, when the subjective magnitude is plotted against physical magnitude, we find that the curves for power functions with different exponents (*N*) have dramatically different shapes as shown in Figure 13.7. If the exponent is less than 1 (such as visual brightness for which $N = 0.33$) the curves are concave downward, which means that as the stimulus becomes more intense, greater stimulus changes are needed to produce the same degree of perceptual change. The result is that the sensation grows more slowly than stimulus intensity and in this respect, it is similar to Fechner's position, that is, a wide range of stimulus intensities is compressed into a small span of subjective values. When exponents are greater than 1 (such as for electric shock for which $N = 3.5$), the curve becomes concave upward which means that as stimuli become more intense, the same physical stimulus change produces an even bigger perceptual change than at lower stimulus intensities. The result is that a small range of physical intensities is expanded into a wider range of subjective values. Here sensation grows more rapidly than stimulus intensity. When exponent (*N*) is exactly 1 as we find in case of apparent length of lines, the function that relates subjective and physical magnitude is a straight line and there is neither compression nor expansion.

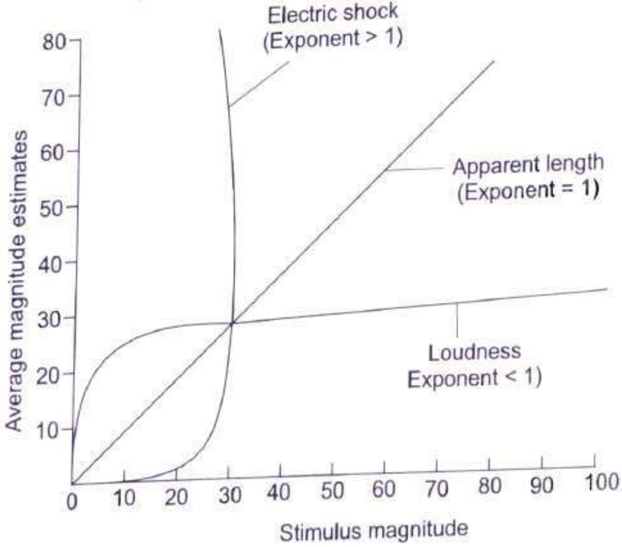

Fig. 13.7

If one is asked to judge the relative merits of Fechner's law and Stevens' law, one will lean in favour of Stevens' law, yet the issue remains unsettled. In fact, Stevens extended Fechner's effort and demonstrated that it was possible for the subjects to deal with sensations head-on and scale them directly.

Here the primary assumption is that categories are in correct rank order and that their boundary lines are stable, except for sampling errors. It is also assumed that distribution of responses to a stimulus is normal on the psychological continuum. There is also an implicit assumption that momentary judgements (placement in categories) are perfectly correlated with momentary responses (psychological attributes).

NEWER PSYCHOPHYSICAL METHODS

Apart from these major psychophysical methods, there are other such methods which have also been used in scaling process. Among them two methods, namely, *method of category scaling* and *method of magnitude estimation* deserve special attention.

Method of Category Scaling

This is a psychophysical scaling method in which stimuli are grouped into a predetermined number of categories on the basis of their perceived strength along certain dimension. In fact, this is one of the direct methods of scaling whose importance was first shown by Sanford. He asked his subjects to judge a number of envelopes, each of which contained different weights, and sort them into five categories. Category 1 was meant for the lightest weight and Category 5 for the heaviest, and the remaining categories were distributed in such a way that intervals between the category boundaries would be subjectively equal. As a consequence, the difference in sensation between the upper and the lower boundaries of Category 1 would be the same as those of Category 2. Thus all categories would be of the same size. That is why this method is also known as *equal-interval scaling*. Since Sanford, category scaling has been studied in greater detail, particularly by Stevens & Galanter (1957), who frequently used a scale of 7 categories rather than 5 categories.

After the subjects have finished the sortings, the mean ratings for each stimulus across subjects are calculated. Such mean ratings definitely reflect an interval scale. Thus absolute level of rated intensity of each stimulus is interpretable, as are the relative distances between the mean ratings for different stimuli. This is one of the major advantages of category scaling. However, there are some disadvantages, too. Here responses are linked to a few labels or categories only. Not only this, in this scaling method, the stimuli that are similar but still discriminably different from one another may be grouped into the same category.

Method of Magnitude Estimation

This method was developed by Stevens (1956). It was originally introduced as a psychophysical method to be applied to the physical stimuli but it has also been successfully applied for scaling psychological stimulus dimensions. This represents the most direct approach to scaling. In this method, the subject is presented with a standard stimulus that serves as the starting point for his judgements. The experimenter also defines the subjective numerical value of the standard stimulus called as *modulus*. The subject's task is to assign each presented stimulus a number that reflects the apparent magnitude of that stimulus relative to standard stimulus. For example, in a typical magnitude estimation experiment the experimenter may wish to scale the loudness of different tones. For this, he would present first, a tone of certain loudness to the subject who would be instructed to call the loudness of this tone as 100 (modulus). After that, the subject would be required to assign all succeeding tones a number relative to that standard. If the next tone was judged twice as loud, he would assign it 200 and if it was judged half as loud, it would be assigned 50. Obviously, this is a very direct way of measuring sensation. It is also clear that the very nature of the task, that is, judging stimulus to be *n* times above or below the standard, implies that the resultant scale will be a ratio scale. Magnitude estimation method was the base of Steven's Power Law. This demonstrates the heuristic value of the method.

Recently, a variation of magnitude estimation method has been developed where the subjects are not required to assign numbers to each stimulus. In this modified form, the subject is required to judge stimuli on the dimension by producing responses on the second dimension. This new version is named as *cross-modality matching*. This may be explained through an example. Subjects may be asked to judge the length of line by varying the loudness of a tone. The louder the tone, the longer the line they judged to be.

PSYCHOLOGICAL SCALING METHODS

Under this head, three important scaling methods, namely, the *method of rank order*, the *method of paired comparison* and the *method of successive categories* will be discussed.

Method of Rank Order

The method of rank order, also known as 'Order of merit method', is one of the simplest techniques for psychological scaling of objects, persons, events and other items. The method of rank order was developed mainly as a sequel to the extensive works of Cattell (1903) and Spearman (1904) who demonstrated how to use rank in the computation of correlation. In this method, all stimuli are presented simultaneously to the subject who is requested to rank them in order from high to low. Conventionally, the rank of 1 indicates the best and the higher rank numbers indicate the inferior position of the stimulus being ranked. The method of rank order is most suited when the number of stimuli or objects to be ranked is small. For the purpose of ranking, all stimuli are presented either physically or symbolically and therefore, no problem regarding the sequence of presentation of objects arises. This is one of the main advantages of the rank order method over the paired comparison method where all stimuli are usually not presented for simultaneous observation.

The scale resulting from the method of rank order is known as an 'ordinal scale'. When ranking has been done, the obtained data may be subjected to either elaborate computational procedures as described by Guilford (1954) or they may be subjected to some simpler treatments, which are common in many research works. Here, we shall limit our discussion to the simpler treatments of rank data. Suppose 10 actors were ranked by 10 judges who have to rank them in terms of how well they liked their acting. Rank 1 indicated the most liked actor and rank 10 indicated the least liked actor. Applying suitable statistical techniques, we can measure the amount of agreement or disagreement among the judges. The data (hypothetical) are given in Table 13.8. From these rankings, a scale can be derived by converting the ranks into choice scores and then to p values and finally to z values. If a judge gives rank 1 to a certain actor, it means that he prefers (or chooses) him to other actors. Likewise, rank 2 indicates that this actor is preferred to the remaining actors, and so on. Thus ranks given by the judge indicate his choice score, C. In general, C can be estimated by the following formula:

$$C = n - r \tag{13.9}$$

where, C = choice score; n = number of objects ranked; and r = rank assigned.

Suppose an object is assigned rank 3 (the total number of objects being 10), its choice score $= 10 - 3 = 7$. Since Equation 13.9 holds true for all the ranks assigned by the judges to the same object, mean rank, M_r and mean choice score, M_c for that object can easily be determined by the equation given below:

$$M_r = \frac{\Sigma r}{N} \tag{13.10}$$

where, M_r = Mean rank of an object;

Σr = Sum of ranks assigned by all judges to an object;

N = Number of ranks (or judges).

$$M_c = n - Mr \tag{13.11}$$

where, M_c = Mean choice score;

n = Number of objects ranked;

M_r = Mean rank.

Subsequently, the value of M_c is converted into p value as under:

$$p = \frac{M_c}{n-1} \tag{13.12}$$

where the symbols are defined as usual. Each p value is converted into z value with the help of a table (Woodworth and Schlosberg 1954, 206). There are certain *checks* on the computations

nod', is one of the simplest
d other items. The method of
works of Cattell (1903) and
utation of correlation. In this
is requested to rank them in
ne best and the higher rank
The method of rank order is
is small. For the purpose of
and therefore, no problem
f the main advantages of the
all stimuli are usually not

s an 'ordinal scale'. When
er elaborate computational
to some simpler treatments,
ur discussion to the simpler
s who have to rank them in
st liked actor and rank 10
ques, we can measure the
(hypothetical) are given in
ting the ranks into choice
ank 1 to a certain actor, it
2 indicates that this actor is
judge indicate his choice

(13.9)

signed.
being 10), its choice score
by the judges to the same
asily be determined by the

(13.10)

(13.11)

(13.12)

value with the help of a
cks on the computations

---

shown above. The average M_r must be equal to $(n+1)/2$. In Table 13.8, the average value of M_r is 5.5 which becomes equal to $(10+1)/2 = 5.5$. Likewise, the average M_c must be equal to $(n-1)/2$. The average value of M_c is 4.5 which becomes equal to $(10-1)/2 = 4.5$. Similarly, the average p must be equal to 1/2 or 0.5 which is a case in Table 13.8.

Table 13.8 Ranks reduced to choice scores and z scores

Judges	Actors									
	A	B	C	D	E	F	G	H	I	J
1	1	2	3	5	4	6	8	7	10	9
2	2	4	3	1	5	7	6	8	9	10
3	1	3	4	2	6	7	8	9	5	10
4	1	2	3	4	6	7	5	8	9	10
5	1	2	3	4	5	6	7	8	10	9
6	1	2	4	3	5	6	8	7	9	10
7	1	3	4	2	5	6	8	7	9	10
8	1	4	2	3	5	6	8	7	10	9
9	2	3	1	4	5	6	7	8	9	10
10	1	2	3	4	6	5	7	8	9	10
Sum =	12	27	30	32	52	62	72	77	89	97
$M_r =$	1.2	2.7	3.0	3.2	5.2	6.2	7.2	7.7	8.9	9.7
$M_c = n - M_r =$	8.8	7.3	7.0	6.8	4.8	3.8	2.8	2.3	1.1	0.3
$p = M_c(n-1) =$	0.98	0.81	0.78	0.76	0.53	0.42	0.31	0.26	0.12	0.03
$z =$	+2.05	+0.88	+0.77	+0.71	+0.08	−0.20	−0.50	−0.64	−1.18	−1.88

Checks: The average $M_r = (n+1)/2$; the average $M_c = (n-1)/2$; the average $p = 1/2$ or 0.5.

The method of rank order has several advantages over its closely related rival called the rating scale. *First*, it forces the judges to use all parts of the scale. In a rating scale a judge may easily use only some parts of the scale, putting all objects into those categories. But this is not true in the case of the rank order method. *Second*, since some objects are to be assigned higher ranks and some objects are to be assigned lower ranks, judges are left with no way to commit the error of central tendency, which is common in rating scales. *Third*, the ranking method requires the judges to differentiate every item from every other item. Thus, the method forces the judges to take a decision regarding every item by making a discrimination, which may easily be evaded in rating method.

The rank-order method has one serious *limitation*. It does not provide the quantitative information regarding the element of the characteristic which segregates the ranks. For example: Can we say that the difference in amount of the characteristic segregating ranks 1 and 2 is equal to the amount of characteristic segregating ranks 2 and 3 or 3 and 4? The answer is probably 'No'. It happens because the rank-order method does not account for the differences among the characteristics being ranked.

In an attempt to refine the rank-order method, Guilford (1954) has suggested that the ranks can be transformed into centiles and subsequently, to another convenient transformation called C scales values. One of assumptions in such a transformation is that the trait or attribute being judged through the different items is normally distributed in the samples of items presented for ranking. The C scale is based upon the normalized transformation of ranks with a mean of 5 and SD of 2. When ranks are transformed into C values, they can easily be treated as interval scale values and all statistical analysis suitable to the interval measurement can be carried out.

Method of Successive Categories

The *method of successive categories,* also known as the *method of successive intervals,* is a technique of scaling originally developed by Thurstone and later on published by Saffir (1937) who made a comparative study of the scale values derived by three different methods, namely, method of pair comparisons, method of rank order and method of successive categories. Like the method of equal-appearing intervals, in the method of successive categories each subject is required to make a single judgement for each statement as belonging to one of a limited number of categories, which tend to differ quantitatively along a continuum. Therefore, the method is a convenient one, especially when the number of statements to be scaled is large. The subjects may be asked to sort each statement in any one of the nine categories mentioned below:

1	2	3	4	5	6	7	8	9
A	B	C	D	E	F	G	H	I
Unfavourable				Neutral				Favourable

Varying degrees of unfavourableness are represented by letters A to D and varying degrees of favourableness are represented by letters F to I. E is the neutral category. No assumption of psychological equality of these categories is made and in this sense, the method differs from that of the equal-appearing intervals where it is so assumed. After the various statements have been sorted or rated into the various successive intervals or categories by a group of judges, the data are obtained in the same form as those obtained in case of the equal-appearing intervals. Suppose a group of 400 judges rated or sorted the 35 statements into nine successive intervals according to the degree of favourableness–unfavourableness that each statement is assumed to express. Of these, the obtained data for four items (for illustrative purposes) have been presented in Table 13.9. The data also show the frequencies with which each statement was sorted into a particular category. These frequencies were cumulated from left to right to yield cumulative frequencies (C_f), which were transformed into cumulative proportions (C_p) by multiplying each C_f by the reciprocal of N. On the basis of the distributions of the cumulative proportions of the set of items, the widths of the intervals making up the psychological continuum are determined. Here we assume that the distribution of the cumulative proportions for each item is normal. Subsequently, the scale values of each item are determined by calculating the mean or median of the cumulative proportions. For the purpose of showing the elaborate calculations involved in estimating the scale values, the cumulative proportions are segregated from Table 13.9 and have been presented in Table 13.10.

The proportions for each item given separately for each successive interval may be designated as p_{ij}. In other words, each cell entry of Table 13.10 may be designated as p_{ij} where i indicates the statement and j indicates the upper limit or boundary of any particular category or the jth category. The full meaning of p_{ij} is the proportion of judges who place the statement i below the upper limit of the jth category and $1 - p_{ij}$ indicates the proportion of judges who place the statement i above the upper limit of the jth category. For instance, p_{36} is 0.400 which means that 0.400 of the judgements for item number 3 fall below the upper limit of the 6th interval and

1 − 0.400 = 0.600 of the judgements for item number 3 fall above the upper limit of the 6th interval. The next step is to convert the p_{ij} into the corresponding normal deviates. If p_{ij} is 0.50, its corresponding normal deviate will be 0.0 and the scale value of the statement *i* will fall exactly on

Table 13.9 Successive intervals data showing frequencies, cumulative frequencies and cumulative proportions for items sorted by 400 judges

Statement		Successive intervals								
		Unfavourable								Favourable
		1	2	3	4	5	6	7	8	9
1	f	2	10	40	100	150	46	30	15	8
	C_f	2	12	52	152	302	347	377	392	400
	C_p	0.005	0.030	0.130	0.380	0.755	0.868	0.942	0.980	1.00
2	f	0	0	100	80	60	80	50	20	10
	C_f	0	0	100	180	240	320	370	390	400
	C_p	0	0	0.250	0.450	0.600	0.800	0.925	0.975	1.00
3	f	0	0	0	20	40	100	200	30	10
	C_f	0	0	0	20	60	160	360	390	400
	C_p	0	0	0	0.050	0.150	0.400	0.900	0.975	1.00
4	f	21	15	10	3	40	170	100	31	10
	C_f	21	36	46	49	89	259	359	390	400
	C_p	0.053	0.090	0.115	0.122	0.222	0.648	0.898	0.975	1.00

Table 13.10 Cumulative proportions (p_{ij}) of 4 items sorted by 400 judges by the Method of successive intervals

Statements	Successive intervals								
	1	2	3	4	5	6	7	8	9
1	0.005	0.030	0.130	0.380	0.755	0.868	0.942	0.980	1.00
2	000	000	0.250	0.450	0.600	0.800	0.925	0.975	1.00
3	000	000	000	0.050	0.150	0.400	0.900	0.975	1.00
4	0.053	0.090	0.115	0.122	0.222	0.648	0.898	0.975	1.00

the upper limit of the jth interval (or on the lower limit of the jth interval plus one). If p_{ij} is greater than 0.50, the corresponding normal deviates will be positive and therefore, the upper limit of the jth interval will deviate positively from the scale value of the statement i. On the other hand, if p_{ij} is less than 0.50, the corresponding normal deviate will be negative and therefore, the upper limit of the jth interval will deviate negatively from the scale value of the statement i. Entering the Table of Normal Deviates* with the p_{ij} of Table 13.10, normal deviates for the successive intervals for each statement were found and presented in Table 13.11. In converting the p_{ij} into the corresponding normal deviates, any value of p_{ij} less than 0.02 and greater than 0.98 is ignored because these two proportions are near 0 and 1 respectively, whose normal deviates are infinite. Hence, the upper category, that is, 9, and the lower category, that is, 1 are generally not scalable. The cell entry or normal deviate of Table 13.11 may be designated as z_{ij} in which i stands for a particular statement and j stands for the upper limit of the jth interval. z_{37} means that the upper limit of interval 7 has been expressed as a normal deviate in terms of item 3. The next step is to compute the width of the successive intervals and for this purpose, some kind of averaging is done. If the z matrix (Table 13.11) is complete, that is, if no cell values are indeterminate, which can happen only when no p_{ij} values are greater than 0.98 and less than 0.02, the width of the successive intervals will be the differences in means of the successive intervals or columns in the z matrix. On the other hand, if the z matrix is incomplete (as it is seen in Table 13.11), the width of the successive intervals is estimated by making successive subtraction in pairs in each row of Table 13.11. The resulting width of the interval is designated as w_{ij} , which simply indicates the width of the jth interval provided by statement i. Thus the width of interval 3 provided by item 1 (w_{13}) is equal to $-1.126-(-1.881)=0.755$. Other estimates of the width of interval 3 have similarly been obtained. The procedure is repeated for obtaining the width of each interval. Table 13.12 presents w_{ij} of the 7 categories. The width of the two extreme categories like category 1 and category 9 cannot be estimated by the procedure being described because the lower limit of category 1 and the upper limit of category 9 cannot be determined. All entries in the given category of Table 13.12 are the estimates of the width of the same interval. Subsequently, the arithmetic means of the entries of each column or category are calculated as the estimates of the width of various successive intervals. The arithmetic means may be designated as $\overline{w}$. In Table 13.12, the first row at the bottom indicates the sum of category entries or column entries. The second row indicates the number of entries in each column; the third row indicates the

Table 13.11 Normal deviates (z_{ij}) corresponding to the upper limits or boundaries of the successive intervals

Statements	Successive intervals								
	1	2	3	4	5	6	7	8	9
1	–	–1.881	–1.126	–0.305	+0.706	+1.126	+1.555	+2.054	–
2	–	–	–0.674	–0.126	+0.253	+0.842	+1.405	+2.054	–
3	–	–	–	–1.645	–1.036	–0.253	+1.282	+2.054	–
4	–1.645	–1.341	–1.175	–1.175	+0.772	+0.385	+1.282	+2.054	–

In converting p_{ij} into z_{ij}, the values of p_{ij} were rounded to two decimal places before conversion.

---

* Table of normal deviates can be seen at p. 246 in *Techniques of Attitude Scale Construction* by A L Edwards (published in India by Vakils, Feffer & Simons, Bombay 1969); American publisher is Appleton Century Crofts, USA.

Table 13.12 Estimates of category widths (w_{ij}) on the basis of the differences between the upper limits of intervals

Statements	Successive intervals						
	2–1	3–2	4–3	5–4	6–5	7–6	8–7
1		0.755	0.821	1.011	0.420	0.429	0.499
2			0.548	0.379	0.589	0.593	0.649
3				0.609	0.783	1.535	0.772
4	0.304	0.166	0.000	0.403	1.157	0.897	0.772
Sum =	0.304	0.821	1.369	2.402	2.949	3.454	2.692
N	1	2	3	4	4	4	4
Mean ($\overline{w}$) =	0.304	0.410	0.456	0.600	0.737	0.886	0.673
Cum ($\overline{w}$) =	0.304	0.714	1.170	1.770	2.507	3.393	4.066

mean or $\overline{W}$ of the categories; and entries in the fourth row indicate cumulating means, which forms the basis for our psychological continuum. Now, when the common psychological continuum has been established, it is easier for the investigator to scale each statement upon the continuum. For finding out the scale value of each statement, the median may be calculated by the following formula:

$$S_i = l + \left(\frac{0.50 - \Sigma p_b}{p_w}\right)\overline{w} \qquad (13.13)$$

where, S_i = scale value of statement i;

l = lower limit of the interval on the psychological continuum upon which the median falls;

Σp_b = the sum of the proportions below the interval in which the median falls;

p_w = proportion within the interval in which the median falls and, therefore, it is equal to the difference between cumulative proportion of the category where median falls and the cumulative proportion of the category falling immediately below the category, which contains the median;

$\overline{w}$ = mean of the interval in which the median falls.

Now, the scale values of the above four statements may be calculated as shown below.

Statements	l	+	$(0.50 - \Sigma p_b)/p_w$	×	$\overline{w}$	=	S_i
1	1.170	+	(0.50 – 0.380)/0.375	×	0.600	=	1.197
2	1.170	+	(0.50 – 0.450)/0.150	×	0.600	=	1.174
3	2.507	+	(0.50 – 0.400)/0.500	×	0.886	=	2.861
4	1.770	+	(0.50 – 0.222)/0.426	×	0.737	=	2.250

The steps involved in calculation of the scale value of statement (or item) No. 1 may be enumerated as shown below.

1. The first step is to determine that interval in which $N/2$ (which is here, 200) falls. A cumulative counting from left to right for item 1 in Table 13.9 shows that $N/2$ falls in the 5th interval or category.

2. The second step is to determine Σp_b, which is the cumulative proportion of the interval immediately below the 5th interval. From Table 13.10, it is read as 0.380.

3. The third step is to determine p_w which is the difference between the cumulative proportion of the interval which contains $N/2$ and the interval immediately below it. From Table 13.10,

$$p_w = 0.755 - 0.380 = 0.375$$

4. The fourth step is to determine the width (or mean) of the interval, which contains $N/2$. From Table 13.12, it is read to be 0.600.

5. The lower limit of interval (l) on the psychological continuum upon which median falls is 1.170 (from cum $\overline{w}$).

6. Finally, all values are substituted in Equation 13.13 and the resulting median becomes the scale value of each statement.

In this way, each statement is scaled by the method of successive categories. Subsequently, the scale values of each statement may be tested for internal consistency if desired by the investigator.

Method of Pair Comparisons

The method of pair (or paired) comparisons is one of the very common methods of scaling. In this method, stimuli are paired and the subject is required to make a comparative judgement by saying which member of each pair is preferred or/possesses more of the trait being scaled. Hence, the method is named as the method of pair comparisons. Before the procedure of the method is considered in detail, it is necessary to examine the *law of comparative judgement*, which is the base of the method of pair comparisons.

Law of Comparative Judgement

Thurstone (1927) formulated a law of comparative judgement, which explained data obtained on the basis of comparative judgements of two stimuli. Each judgement in comparative judgement does not yield a quantitative value on a psychological continuum. But the law of comparative judgement made this possible through a series of computational processes. The law may be defined as a series of assumptions, which relate the proportion of times one stimulus is judged higher on a given attribute than the other stimulus (of the pair) to the scale values and discriminal dispersions of these two stimuli which are repeatedly being compared. The set of assumptions is derived from the following four postulates:

1. The stimulus presented to the subject produces a response, which has some value on the psychological continuum. In Thurstone's terminology, this response is known as the 'discriminal process', which may be designated as S_i. Thurstone (1927) defined the discriminal process as "that process by which the organism identifies, distinguishes, or reacts to stimuli." Thus a discriminal process is a theoretical concept and indicates the subject's reaction or response when asked to make a judgement regarding any one of two stimuli on a given attribute.

2. A given stimulus, if presented repeatedly, does not produce the same discriminal process. Sometimes, the value of the discriminal process may increase and sometimes, it may decrease. The reasons why the changes in the value of the discriminal process of the same stimulus occur are the momentary fluctuations in the subject.

3. If any stimulus is presented to the subject a larger number of times, it will generate a frequency distribution, which is a normal one. Thus, each stimulus generates a normal distribution of discriminal processes. The distribution is illustrated in Figure 13.8.

4. The mean and standard deviation of the normal distribution of the discriminal process generated by a stimulus corresponds to its scale value and the discriminal dispersion respectively.

The discriminal process which is most frequently aroused by a given stimulus is known as the 'modal discriminal process' and may be designated as $\bar{S}_i$ (See Figure 13.8). Since in normal distribution the mean, median and mode have the same value, the modal discriminal process, being the mean of the set of discriminal processes aroused by a given stimulus, is taken as the scale value of that stimulus. The scale values (or the modal discriminal process) and the discriminal processes for the different stimuli will not necessarily be the same. This point has been illustrated in Figure 13.9.

The distribution of discriminal processes obtained from four stimuli has been shown in Figure 13.9. The scale value of each stimulus is its modal discriminal process. Thus, the scale value of stimulus 1 is $\bar{S}_1$, the scale value of stimulus 2 is $\bar{S}_2$, and so on.

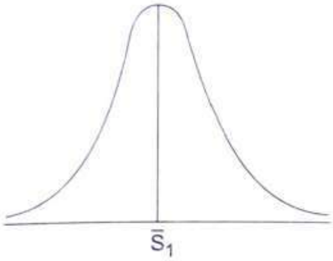

Fig. 13.8 Normal distribution of discriminal process generated by stimulus

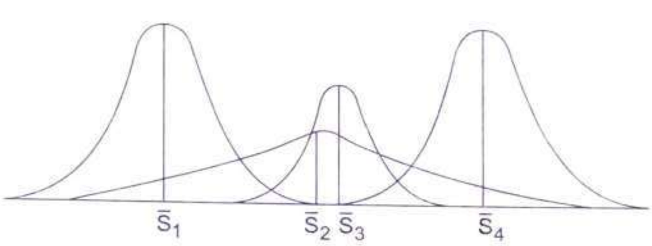

Fig. 13.9 Distribution of discriminal process generated by the four stimuli on the same psychological continuum

Since the subject cannot directly tell the value of the discriminal process on the given psychological continuum (that is, heavier, brighter, louder), no frequency distribution of discriminal processes can be directly obtained. Hence, the scaling of stimulus is done indirectly through a set of equations which relates the judgements made by the subjects to the scale values and discriminal dispersions on the psychological continuum. The law of comparative judgement does this job well. Let *i* and *j* stand for two statements (or stimuli) on a psychological continuum. Suppose that a group of subjects has been asked to judge as to which of these two statements is more favourable. On each pair presentation to each individual, *i* and *j* will elicit a separate discriminal process. Let the discriminal process aroused by i be designated as S_i and the discriminal process aroused by *j* be designated as S_j. The difference between S_i and S_j is known as the 'discriminal difference'. If these two statements are presented to each member of the group, the discriminal differences themselves are likely to form a normal distribution on the psychological continuum. As we know from the elementary principles of statistics, the difference between two normally distributed variables is also normally distributed with a standard deviation. Thus

$$\sigma_{i-j} = \sqrt{\sigma_i^2 + \sigma_j^2 - 2r_{ij}\sigma_i\sigma_j} \qquad (13.14)$$

where, σ_{i-j} = Standard deviation of the differences between the discriminal processes of *i* and the discriminal processes of *j*, that is, SD of the discriminal differences ($S_i - S_j$);

σ_i = Standard deviation of the discriminal processes generated by *i*, that is, SD of S_i;

σ_j = Standard deviation of the discriminal processes generated by *j*, that is, SD of S_j;

r_{ij} = Correlation between S_i and S_j.

Thurstone has presented *five cases* of the law of comparative judgement. His Case I is a complete form of the law based upon Equation 13.14 and the normal deviates, z_{ij} corresponding to the proportions of comparative judgements, (that is, P_{ij}). Thus P_{ij} is the proportion of times i is judged greater or higher than j. When we wish to express the proportion of times j is judged greater than i, the subscript becomes P_{ji}. When the comparative judgements of i and j have been obtained by a group of judges, we have frequencies, which show how many subjects judged i greater than j. Accordingly, f_{ij} may be written to indicate those frequencies. When f_{ij} is multiplied by the reciprocal of N, we get the value of P_{ij} which, in turn, can be expressed as a unit of normal deviates z_{ij} with the help of the Table of normal deviates. The equation for the Case I of the law of comparative judgement in its complete form is based upon Equation 13.14 and may be written as under:

$$\bar{S}_i - \bar{S}_j = z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2 - 2r_{ij}\sigma_i\sigma_j} \tag{13.15}$$

where, $\bar{S}_i$ and $\bar{S}_j$ indicate the modal discriminal process or scale values of statements i and j respectively. The remaining symbols are defined as usual. Equation 13.15 applies to the repeated judgements of the single subject. Thurstone's Case II of the law of comparative judgement is essentially parallel to his Case I and requires the same kind of complete data as in Equation 13.15. His Case II is meant for a situation where several subjects make a single judgement in each possible pair and Equation 13.15 applies to this situation, too. The complete solution of Equation 13.15 is not possible because we have more unknowns than knowns. Hence, three approximations to Equation 13.15, which simplify the assumptions associated with it have been proposed. These are known as Case III, IV, and V of the law of comparative judgement.

Suppose we have five statements expressing attitudes towards birth control. Now these statements are presented in all possible pairs to a group of 200 subjects. The maximum number of pairs from five statements will be given by $\frac{N(N-1)}{2} = \frac{5(5-1)}{2} = \frac{5 \times 4}{2} = 10$. For five statements we will have the 10 equations given by Eq. 13.15 and these 10 equations will have 20 unknowns, that is, 5 scale values, 5 standard deviations and 10 intercorrelations and 5 knowns, that is, the values of z_{ij} calculated on the basis of p_{ij} with the help of the Table 13.11 of the normal deviates. In order to solve the problems faced by the unknown intercorrelations, Thurstone assumed that the intercorrelations are zero, that is, no correlation exists between the responses to any pair of statements. This assumption is more defensible where the stimuli are not clearly identifiable. As a consequence, the law reduces to the following form (because the last part is dropped):

$$\bar{S}_i - \bar{S}_j = z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2} \tag{13.16}$$

Equation 13.16 illustrates Case III of the law of comparative judgement. Still, in the above examples we will be having more unknowns (5 scale values and 5 standard deviations) than knowns. To solve this problem faced by the discriminal dispersions or standard deviations, Thurstone assumed that the discriminal dispersions are nearly equal. Then, the law reduces to

$$\bar{S}_i - \bar{S}_j = 0.707 \ z_{ij}(\sigma_i + \sigma_j) \tag{13.17}$$

Equation 13.17 illustrates the Case IV of the law of comparative judgement. Discriminal dispersions will be equal when stimuli are equally easy to place on a scale. Now Thurstone made one additional assumption to simplify his law further. He assumed that all discriminal dispersions are equal, that is, $\sigma_i = \sigma_j$ and then the law reduces to

$$\bar{S}_i - \bar{S}_j = z_{ij}\sigma_i \sqrt{2} \tag{13.18}$$

in which σ_i stands for both discriminal dispersions. Taking $\sigma_i\sqrt{2}$ to be the common unit of measurement of scale separations of the various pairs of statements or stimuli and letting this common unit of measurement be equal to unity or 1, the law finally reduces to

$$\bar{S}_i - \bar{S}_j = z_{ij} \tag{13.19}$$

$$\text{or} \quad z_{ij} = \bar{S}_i - \bar{S}_j$$

Equation 13.19 illustrates Case V of the law of comparative judgement. It is theoretically an equal interval scale. Still, we are left with 5 unknowns, that is, 5 scale values. The steps involved in the calculation of scale values will be discussed in the section below.

Experimental procedure of the method of paired comparisons

The law of comparative judgement discussed in the previous section, requires that each stimulus should be compared with every other stimulus a large number of times so that the proportions of times stimulus *i* has been judged greater or higher than stimulus *j* is clearly known. The method for having some empirical estimates from these proportions is technically known as the *method of paired comparison.*

Originally, the method of paired comparison was introduced by Cohn in 1894 in his study of colour preferences of subjects. Subsequently, it was further developed by Thurstone. As a matter of fact, the method of paired comparison is nothing but an extension of the method of constant stimulus difference having two-category responses. In the method of constant stimulus difference, each time the two stimuli are presented to the subject in which there is one standard stimulus and one comparable stimulus. Thus, each time each stimulus is compared with a single standard. In the method of paired comparisons each stimulus is compared with every other stimulus and therefore, each stimulus serves as a standard stimulus in turn. When the subject is presented with the two stimuli in pairs, his job is to tell which stimulus of the pair is greater (heavier, more favourable, louder) on the psychological continuum (or the attribute being scaled). No equality judgement is allowed and therefore, the subject must identify one of the members as being greater than the other. Even if the subject finds some difficulty in making a comparative judgement, he is encouraged to give his judgement in favour of any one member of the pair because in the method of pair comparison, there is no *provision* for analysis of the responses showing failure to select one member of the pair. Ordinarily, a stimulus is not compared with itself and it is assumed that if such judgements are obtained, their proportion will be 0.5 or their frequency will be half of the total number of subjects/judgements. It should be, however, noted that there is nothing in the method itself which restricts such judgement.

In the method of paired comparison, some biasing effects in the form of space error, time error, fatigue effect, practice effect, may occur and it is, therefore, essential that such effects be adequately controlled. In his law of comparative judgement, Thurstone has made no provision for controlling these effects and hence, it deserves more attention. Perhaps, the most common method of controlling space error is to present one member of the pair right (or above, etc.) for half of the subjects or trials and the other member left for the other half of the subjects or trials. Similarly, time error can be controlled by presenting one member of the pair first for half the time and by presenting the same member of the pair second for the remaining half time. Likewise, by reversing the order of the presentations of pairs for half of the subjects or trials, the practice effect or fatigue effect may be controlled. Attempt should also be made to keep pairs having one member common maximally separated in the order of presentation.

Ordinarily, there are two criteria for using the method of paired comparison. *First,* the number of stimuli to be scaled should be small. If the number is large, the method should preferably not be used because it would tax the motivation and interests of the subject. *Second,* where the purpose is to obtain an interval scaling of the psychological dimension, the method is preferred. In order to obtain data suitable for being analyzed by the method of paired comparison, it is essential that a large number of comparisons be made for each pair of stimuli. For this purpose, the situation must be in accord with any of following three conditions:

1. A single subject must have judged each pair a large number of times.
2. Many subjects must have judged each pair only once.
3. Many subjects must have judged each pair several times.

Now, we can return to the details of the empirical estimates of the scale values to be made by the method of paired comparison. In the previous examples, 5 statements relating to attitude towards birth control were judged in all possible pairs by 200 subjects so that each subject made $[5(5-1)]/2 = 10$ comparative judgements. The raw data showing the number of times one statement of the pair was judged more favourable than the other statement have been arranged in the $N \times N$ matrix, called the F matrix, which gives the frequency with which each column statement or stimulus was judged more favourable than the row statement or stimulus (see Table 13.13).

Table 13.13 The F matrix for 5 statements judged by 200 subjects

Statements	1	2	3	4	5
1	100	150	170	160	170
2	50	100	120	150	180
3	30	80	100	120	180
4	40	50	80	100	110
5	30	20	20	90	100

In Table 13.13, diagonal entries are assumed to be equal to $N/2$ because no comparison of a statement with itself is made. In this example, each subject made one comparative judgement for each pair only once. Hence, for each pair there were 200 comparative judgements, which is equal to the number of subjects. On the basis of the F matrix of Table 13.13, the P matrix shown in Table 13.14 is prepared. Each entry in Table 13.13 has been multiplied by the reciprocal of N. The N is 200 and its reciprocal is 0.005.

Table 13.14 *P* matrix prepared on the basis of *F* matrix, that is, Table 13.16

Statements	1	2	3	4	5
1	0.500	0.750	0.850	0.800	0.850
2	0.250	0.500	0.600	0.750	0.900
3	0.150	0.400	0.500	0.600	0.900
4	0.200	0.250	0.400	0.500	0.550
5	0.150	0.100	0.100	0.450	0.500
Sums:	0.750	1.500	1.950	2.600	3.200

The entries shown in the P matrix indicate the proportion of times the column stimulus is judged favourable to the row stimulus and may be abbreviated as p_{ij}. Subsequently, each column in the P matrix has been summed ignoring the diagonal entries. It is desirable to rearrange the stimuli or statements in terms of the increasing order of rank of column sums of the P matrix so that the smallest column sum must be at the left and the largest column sum must be at the right of the p matrix. While making this rearrangement, the corresponding pairs in row and column should both be interchanged. For example, if one interchanges the column entries for statement 3 and statement 2, the corresponding row entries for statements 3 and 2 should also be interchanged. We find in Table 13.14 that no rearrangement of rows and column entries is needed because the sum of each column has already been obtained in terms of increasing order of ranks. Now, with the help of the Table of Normal Deviates, we estimate the unit normal deviate, z_{ij} corresponding to p_{ij} entries of Table 13.14. Table 13.15 shows the value of z_{ij} and is known as the Z matrix. An inspection of the Z matrix reveals that if p_{ij} is higher than 0.500, it gets a positive sign and if p_{ij} is lower than 0.500, it receives a negative sign.

Table 13.15 *Z* matrix prepared on the basis of the *P* matrix, that is, Table 13.14

Statements	1	2	3	4	5
1	000	0.674	1.036	0.842	1.036
2	– 0.674	000	0.253	0.674	1.282
3	– 1.036	– 0.253	000	0.253	1.282
4	– 0.842	– 0.674	– 0.253	000	0.126
5	–1.036	–1.282	–1.282	– 0.126	000
Sums =	–3.588	–1.535	– 0.246	1.643	3.726 $\Sigma = 0$
N =	5	5	5	5	5
Mean =	– 0.718	– 0.307	– 0.049	0.329	0.745 $\Sigma = 0$
Mean + 0.718 =	000	0.411	0.669	1.047	1.463 $\Sigma = 3.59$

A p_{ij} of 0.500 is always equal to z_{ij} of 000. It becomes also obvious from the Z matrix that all the values of z_{ij} above the diagonal entries of zero (that is, the upper right portion of the matrix) are numerically identical with the values of z_{ij} below the diagonal entries, that is, the lower-left portion of the matrix except for the signs. Now if the entries in column 2 of Table 13.15 are subtracted from the entries in column 1, we would have, according to Equation 13.19, that is, Case V of Thurstone's law of comparative judgement, the equations given below:

$$z_{21} - z_{11} = (\bar{S}_2 - \bar{S}_1) - (\bar{S}_1 - \bar{S}_1) = \bar{S}_2 - \bar{S}_1 = 0.674 - (000) = 0.674$$

$$z_{22} - z_{12} = (\bar{S}_2 - \bar{S}_2) - (\bar{S}_1 - \bar{S}_2) = \bar{S}_2 - \bar{S}_1 = 000 - (-0.674) = 0.674$$

$$z_{23} - z_{13} = (\bar{S}_2 - \bar{S}_3) - (\bar{S}_1 - \bar{S}_3) = \bar{S}_2 - \bar{S}_1 = -0.253 - (-1.036) = 0.783$$

$$z_{24} - z_{14} = (\bar{S}_2 - \bar{S}_4) - (\bar{S}_1 - \bar{S}_4) = \bar{S}_2 - \bar{S}_1 = -0.674 - (-0.842) = 0.168$$

$$z_{25} - z_{15} = (\bar{S}_2 - \bar{S}_5) - (\bar{S}_1 - \bar{S}_5) = \bar{S}_2 - \bar{S}_1 = -0.1282 - (-0.1.036) = -0.246$$

Or, in general,

$$z_{2j} - z_{1j} = (\bar{S}_2 - \bar{S}_j) - (\bar{S}_1 - \bar{S}_j) = \bar{S}_2 - \bar{S}_1 \tag{13.20}$$

Similar equations can be written for the entries in column 3 minus the entries in column 2, for the entries in column 4 minus the entries in column 3, for entries in column 5 minus the entries in column 4, and so on.

When the values shown above are summed together and divided by N, we get 0.411, which is equal to the difference between the mean of the entries in column 2 and mean of the entries in column 1. Since the difference between the means is equal to the mean of the differences, we can get the same results by summing the entries in each column and dividing by N and finding the means. This has been done for each column. Means, thus obtained, become the scale values of the statements or stimuli. In the first row at the bottom of Table 13.15 are the sum of each column. The sum of these sums should be zero which is a check upon the calculation. The second row at the bottom represents the number of entries in each column. The third row represents the means of each column. These means become the scale values for the concerned statements. The sum of mean should be zero, which is taken as a check upon the accuracy of the calculation. Some of the scale values are negative and some are positive. For convenience one often wishes to avoid negative signs in calculation. One way to get rid of a negative sign is to assign zero value to the lowest statement or stimulus. For this purpose, a constant positive value, equivalent to the scale value of the lowest stimulus, is added to all the scale values. Here, a constant of +0.718 is added to the means of all the columns, which make the scale value of the lowest statement zero and the scale values of all other statements positive. Since the mean of the scale values of the statements on the given psychological continuum is arbitrary, the addition of a constant like above does not change the distance between any of the scale values nor their location on the psychological continuum. As a check on accuracy of calculation, the sum of means after adding the constant should be equal to *n* times the constant added. Here, $+0.718 \times 5 = 3.590$, which equals the sum of means after adding the constant, that is, 0.718.

One should note carefully that when the Z matrix is incomplete, that is, some cells are vacant, the scale values cannot be estimated by the procedure described above. The vacancies in the cells of the Z matrix occur when the corresponding p_{ij} are at extreme. z_{ij} becomes indeterminate when p_{ij} is either near 1.00 or near 0. Such extreme proportions are likely to introduce unreliability into the estimate of scale separations. Edwards (1957) has therefore suggested that all these comparative judgements be ignored whose p_{ij} is greater than 0.98 or less than 0.02. The z_{ij} for these two values of p_{ij} is +2.054 and –2.054 respectively. Guilford (1954) has recommended that z_{ij} that are more extreme than +2.0 and –2.0 be disregarded for which the corresponding p_{ij} is 0.977 and –0.023 respectively.

Having determined the scale values of the statement or stimulus by the method of paired comparisons, we can next apply a test of internal consistency of scale values. The internal consistency check comprises comparing the empirical data obtained in proportions p_{ij} with those to be obtained in terms of our derived scale values. The smaller the difference between the empirical proportions p_{ij} and the expected proportions p_{ij}, the higher is the consistency of the scale values. Ideally, if the difference is zero, this indicates perfect consistency. But this is rarely achieved. The first step in applying the test of internal consistency is to obtain the theoretical normal deviates z_{ij} for scale separations or distances of the statements. For this we set up a table like Table 13.16 in which at the top and on the left side, the scale values of the statements are shown. Subsequently, the scale values written on the left side are subtracted columnwise from the scale values written at the top. The value thus obtained becomes the value for z'_{ij}. For example, in Table 13.16, when the scale value of stimulus 2 (the second row) is subtracted from the scale

Table 13.16 Theoretical normal deviates z'_{ij} corresponding to the scale separations of the statements shown in Table 13.18

Statements	Scale values	1	2	3	4	5
		000	0.411	0.669	1.047	1.463
1	000	–				
2	0.411	−0.411	–			
3	0.669	−0.669	−0.258	–		
4	1.047	−1.047	−0.636	−0.378	–	
5	1.463	−1.463	−1.052	−0.794	−0.416	–

value of stimulus 1 (the first column), we get 000–0.411 = –0.411, and so on. The theoretical normal deviates z'_{ij} are obtained for entries below the diagonal entries of Table 13.16. Now by entering in the Table of Normal Deviates we obtain the corresponding theoretical proportions p'_{ij}. For example, for z'_{ij} of –0.411 we get the p'_{ij} value of 0.341; for z'_{ij} of 0.669, we get $p_{ij}{}'$ of 0.252, and so on (cf Table 13.17). Having determined $p_{ij}{}'$ for each entry below the diagonal, we next determine the difference between the empirical p_{ij} of Table 13.14 and the corresponding expected theoretical proportions p'_{ij} of Table 13.17. For this purpose, we subtract each entry of Table 13.17 from the corresponding entries of Table 13.14 (that is, $p_{ij} - p'_{ij}$). Table 13.21 shows the discrepancies between p_{ij} and p'_{ij}. Subsequently, each column in Table 13.18 is summed ignoring the signs. The resulting absolute values are divided by $N(N-1)/2$ or the number of discrepancies. In this way, we get the average error or discrepancies.

Table 13.17 Theoretical proportions p'_{ij} corresponding to the theoretical normal deviates z'_{ij} of Table 13.19

Statements	1	2	3	4	5
1	–				
2	0.341	–			
3	0.252	0.398	–		
4	1.148	0.262	0.353	–	
5	1.126	0.146	0.214	0.339	–

Table 13.18 Discrepancies between theoretical proportions p'_{ij} and empirical proportions p_{ij}

Statements	1	2	3	4	5
1	–				
2	– 0.091	–			
3	– 0.102	0.002	–		
4	– 0.052	0.012	0.047	–	
5	– 0.024	0.046	0.114	0.111	–
Sums =	0.269	0.060	0.161	0.111	–

Thus,

$$\text{Average error or discrepancy} = \frac{\Sigma(p_{ij} - p'_{ij})}{N(N-1)/2} \qquad (13.21)$$

$$= \frac{0.601}{10} = 0.0601 = 0.06$$

This value of average error for 5 statements is in accord with the values commonly obtained by the different investigators. Hevner (1930), for example, reported an average discrepancy or error of 0.024 for 20 stimuli when scaled by the method of paired comparison.

The method of paired comparison has wider applicability. In general this method is readily applied whenever stimuli can be presented in pairs, either simultaneously or in succession. The method has been widely used in determination of affective values as well as of aesthetic values. The favourite stimuli for this method are colours, designs, names of personalities, names of countries, etc. Some psychologists have also evaluated the opinions on such questions as prohibition, war, religion and the like by this method. In view of its wider applicability, it is hoped that it will soon replace the rating scales where more exacting practical or experimental work needs to be done.

METHODS OF ATTITUDE SCALES OR OPINIONNAIRES

Although the method of scaling discussed in the previous section can be applied to the construction of attitude scales, a separate section for attitude scales is essential because the previous scaling techniques are relatively less common in attitude scaling. The measurement of attitude is necessarily done indirectly. Attitude can be measured only on the basis of inferences drawn from verbal statements regarding belief, feeling and tendency to act towards the object or person. Attitude scales (also known as opinionnaires), which usually consist of a large number of statements towards objects of attitudes, are one such indirect measure. Here we shall discuss the three most common and frequently used techniques in construction of attitude scales, namely, the *method of equal-appearing intervals,* the *method of summated ratings* and the *method of cumulative scaling.*

Method of Equal-Appearing Intervals

We have seen in the previous section that the method of paired comparison is not well suited to the situation where the number of statements to be scaled is large, probably because subjects do not have the patience to make a large number of comparative judgements. In such a situation, the solution is to scale the statements through the method of equal-appearing intervals where each subject is required to make only one comparative judgement for each statement. In the history of development of techniques for the construction of attitude scales, Thurstone (1929), Thurstone & Chave (1929), Thurstone (1931) developed a method known as the method of equal-appearing intervals. The method, as used by Thurstone and Chave, is given below.

A large number of statements, both favourable and unfavourable, towards the object of attitudes, are collected from various sources. The number of items usually ranges from 100 to 200. Each statement is printed on a separate card and subjects are asked to sort these printed statements into a number of intervals. Along with the statements, each subject is given 11 cards on which letters A to K are written. These cards are arranged before the subjects in a manner that A is kept at the extreme left and K is kept at the extreme right. A indicates the most unfavourable interval and K represents the most favourable interval. The middle category is designated by the letter F, which is the neutral category. The cards lettered from G to K represent various degrees of favourableness and the cards lettered from D to A represent various degrees of unfavourableness as illustrated below:

A	B	C	D	E	F	G	H	I	J	K

Unfavourable Neutral Favourable

Thurstone and Chave defined only the two extremes and the middle category (of the 11 intervals) on the ground that the undefined intervals between successive cards would represent equal-appearing intervals for all the subjects. The subjects are requested to sort the given statements in terms of 11 intervals represented by the 11 cards. Ordinarily, there is no time limit for sorting. But Thurstone and Chave reported that subjects took 45 minutes in sorting 130 statements into 11 intervals. They made the following assumptions in this method:

1. The intervals into which the statements are sorted or rated are equal.

2. The attitude of the subjects does not influence the sorting of the statements into the various intervals. In other words, subjects having favourable attitudes and those having unfavourable attitudes would do the sorting in a similar manner. Thus the scale values of the statement are independent of the attitude of the judges.

There is no fixed number of subjects to be engaged in the sorting work. However, Thurstone and Chave used 300 subjects in sorting 130 statements. Other investigators like Edwards & Kenney (1946) and Ferguson (1939) have demonstrated that reliable sorting can be done even with a smaller number of subjects. When sorting is over, the next step is to determine the scale value and *Q* value of each statement. For this, the easiest way is to arrange the frequencies of sorting as shown in Table 13.19.

Table 13.19 Frequencies, median and *Q* value of the judgements obtained by the method of equal-appearing intervals (N = 400)

Statement		Sorting intervals											Scale value	Q value
		A	B	C	D	E	F	G	H	I	J	K		
		1	2	3	4	5	6	7	8	9	10	11		
1	f	2	10	4	10	50	20	200	60	30	8	6	7.02	0.525
	C_f	2	12	16	26	76	96	296	356	386	394	400		
2	f	0	0	10	20	20	100	160	60	110	10	10	6.81	1.38
	C_f	0	0	10	30	50	150	210	270	380	390	400		
3	f	0	20	10	90	100	10	10	20	120	10	10	5.300	2.28
	C_f	0	20	30	120	220	230	240	260	380	390	400		
4	f	4	10	20	20	10	40	82	130	60	20	4	7.61	0.99
	C_f	4	14	34	54	64	104	186	316	376	396	400		

Subsequently, the median and *Q* are calculated for each statement separately. The median becomes the scale value of each statement and *Q* indicates the extent of disagreement among the subjects regarding the degree of attributes possessed by the statement. For illustrative purposes, the details of calculation involved in estimating the median and *Q* value of the first statement are given below.

$$\text{Mdn} = l + \frac{\left(\frac{N}{2} - F\right)}{f_m} i \qquad (13.22)$$

where Mdn = median; l = the lower limit of the interval in which the median falls; F = sum of all frequencies below l or the cumulative frequency of the interval below l; f_m = frequency of the interval that contains the Mdn; and i = width of the interval which is here assumed to be equal to 1.00.

Substituting the values in Equation 13.22 for the first statement, we have

$$\text{Mdn} = 6.5 + \left(\frac{\frac{400}{2} - 96}{200}\right)1$$

$$= 6.5 + \frac{(200-96)1}{200} = 6.5 + \left(\frac{104}{200}\right)1 = 6.5 + 0.52 = 7.02$$

Thus the scale value or median of the first statement is 7.02. For estimating Q, Q_1 and Q_3 are to be calculated first. Q_1 may be calculated by the following equation:

$$Q_1 = l + i\,\frac{(N/4 - \text{cum}.f)}{f_q} \qquad (13.23)$$

where, Q_1 = first quartile or 25th percentile;
l = lower limit of the interval in which $N/4$ falls;
i = width of interval which is here assumed to be 1.00;
$\text{cum}\cdot f$ = cumulative frequency of the interval just below l; and
f_q = the frequency of interval which contains $N/4$.

Likewise, Q_3 or the 75th percentile may be calculated by the following equation:

$$Q_3 = l + i\,\frac{(3N/4 - \text{cum}.f)}{f_q} \qquad (13.24)$$

where, Q_3 = third quartile or 75th percentile;
l = lower limit of the interval in which $3N/4$ falls;
i = width of the interval which is here equal to l;
$\text{cum}\cdot f$ = cumulative frequency of the interval just below l; and
f_q = frequency of the interval which contains $3N/4$

Now, substituting the values in Equation 13.23 and Equation 13.24, we get

$$Q_1 = 6.5 + 1\left(\frac{100-96}{200}\right) = 6.5 + 1\left(\frac{4}{200}\right) = 6.5 + 0.01 = 6.52$$

$$Q_3 = 7.5 + 1\left(\frac{300-296}{60}\right) = 7.5 + 1\left(\frac{4}{60}\right) = 7.5 + 0.067 = 7.567 = 7.57$$

Now, Q may be estimated by the equation

$$Q = \frac{Q_3 - Q_1}{2} \qquad (13.25)$$

$$Q = \frac{7.57 - 6.52}{2} = 0.525$$

Similar procedures have been repeated in the calculation of scale values and Q values for the remaining three items. These values have been written in Table 13.19. As we know, Q is a measure of the spread of the middle 50% of judgements. When the judges are in close agreement with the degree of favourableness or unfavourableness shown by a statement, the value of Q will be small. When the value of Q is large, it is a clear indication that there is something wrong with the concerned statement. Such statements are generally considered to be vague and ambiguous and, therefore, should be either modified or dropped. Guilford (1954, 458) has said, "Items with largest Q value should be omitted first." According to Thurstone & Chave, when the different subjects interpret the meaning of a statement in different ways, the value of Q becomes larger and hence, such items should be dropped.

If one wishes to determine the scale values and the Q values through the graphical method (and not through the method as shown above), the best way to do it is to plot the ogive of the judgements obtained for each item separately. The interval on the psychological continuum can be shown on the x axis and the cumulative per cent frequency C_{f_p} can be shown on the y axis. The cumulative per cent frequency is calculated by the following formula:

$$C_{fp} = C_f \times \frac{100}{N} \qquad (13.26)$$

where, C_{fp} = cumulative frequency per cent; C_f = cumulative frequency; and

N = number of judgements

In terms of Equation 13.26, C_{fp} for the first interval of the first item is equal to $(2 \times 100)/400 = 0.5$; for the second interval it is $(12 \times 100)/400 = 3$, and so on. As a check on calculation, the C_{fp} of the 11th interval should be equal to 100. If three perpendiculars one each from 25th cumulative per cent, 50th cumulative per cent and 75th cumulative per cent are drawn across the curve on the baseline or X axis, we will get three separate values. The value where the perpendicular drawn from the 50th cumulative per cent touches the baseline, will be the scale value or the median. The value where the perpendicular drawn from 25th cumulative per cent touches the baseline will be Q_1, and the value where the perpendicular drawn from 75th cumulative per cent touches the baseline will be Q_3. The one-half distance between Q_3 and Q_1 becomes the Q value.

Besides scale value and Q value, there is one more criterion used frequently in the selection of statements by the method of equal-appearing intervals. This is known as the 'criterion of relevance', which is based upon agreement or disagreement of subjects with the statements. After determining the scale value and Q value for each statement, Thurstone and Chave presented 130 statements to 300 subjects (not for the purpose of measuring their attitude) with the request to check those statements with which they agreed. Subsequently, for the purpose of knowing the internal consistency of the statements, it was determined how many items or statements pass the criterion of relevance. If the subjects who agreed with a particular statement were found to check those statements having widely different scale values, the statement was said to be irrelevant for the attitude being measured and hence, was rejected. On the other hand, if the subjects who agreed with a particular statement were found to check the statements with similar scale values, the statement was said to be a relevant one for the attitude being measured and hence, included in the scale. Finally, about 20 to 22 statements were selected. When the choice for final selection from among statements with approximately equal scale values is to be made, generally those statements are selected whose Q values are small. Persons whose attitude is to be assessed are given a scale consisting of 20 to 22 statements with the request to endorse those items with which they agree. The attitude score of the person is the mean or median of the scale values of all those statements endorsed by him. For example, suppose that the testee or examinee agrees with five statements having scale values 3.2, 3.6, 4.7, 4.9 and 5.8. Then his attitude score will be a sum of these five ratings divided by the number of items, that is 22.2/5 = 4.4 indicating that he is slightly unfavourable towards the attitude object since 6 is the midpoint of 11-point scale. Generally, it is customary on the part of the investigator to construct the parallel form of the scale when it is being constructed by the method of equal-appearing intervals.

The method of equal-appearing intervals has its advantages and disadvantages. Its important advantages are given below.

1. Thurstone scales enable the researcher to differentiate between larger numbers of people regarding their attitudinal position. Here item weights are averaged (median) and this reveals a great variety of attitudinal positions. This makes it possible to make finer distinctions among people according to the attitudes they have.

2. In Thurstone scale, it can be said with increased confidence that items being used have a stronger claim to reliability because they are based upon judges' view who have higher degree of agreement about items used and who eliminate the bad items reflecting little or no agreement.

This method has the following disadvantages:

1. Judges or subjects do not keep the intervals equal. Fransworth (1943) has found evidence in the support of the above fact. As a matter of fact, in the method of equal-appearing intervals there is no way through which this assumption can be tested. Thus one of the assumptions of the method does not stand the rigours of the experimental test.

2. It is also said that the attitude of the subjects or judges tends to influence the sorting of the statements into the intervals. In other words, the scale values of the statement are not independent of the attitudes of the judges who do the sorting.

3. Thurstone and Chave have provided no objective basis for selecting the most discriminating items from among items having approximately the same scale values. It may be possible that items having approximately same scale values differ in their discriminatory power.

4. The subjects may do the sorting work carelessly and with least interest. In such a situation, the interpretation of the scale values may be a difficult task. Thurstone and Chave, have, however provided a technique to detect careless judgements and accordingly, they can be eliminated. They have pointed out that if any subject sorts more than 30 statements in any one of the 11 intervals, the judgements of that subject may be rejected on the ground that he has done sorting either carelessly or has misunderstood the instruction.

Method of Summated Ratings

Likert (1932) developed a different method for the construction of the attitude scale. His method, named by Bird (1940, 159) as the method of summated ratings, is a simpler method than that of Thurstone's equal-appearing intervals method. The main steps involved in Likert's method may be summarized as mentioned below:

1. A large number of multiple-choice-type statements usually with five alternatives such as *strongly agree, agree, undecided, disagree* and *strongly disagree* concerning the object of attitude are collected by the investigator. Two examples of items intended to measure attitude towards nationalization are given below:

Nationalization improves the economy of the country.

Strongly agree	Agree	Undecided	Disagree	Strongly disagree
(5)	(4)	(3)	(2)	(1)

Nationalization introduces a feeling of carelessness among employees.

Strongly agree	Agree	Undecided	Disagree	Strongly disagree
(1)	(2)	(3)	(4)	(5)

2. Such statements are administered to a group of subjects who respond to each item by indicating which of the given five alternatives they agree with.

3. Every responded item is scored with different weights. The weight ranges from 5 to 1. For favourable statements a weight of 5 is given to 'Strongly agree', 4 to 'Agree', 3 to 'Undecided', 2 to 'Disagree', and 1 to 'Strongly disagree' (as shown in the first illustrative example), and for the unfavourable statement the order of weights to be given is reversed so that 'Strongly agree' receives I and 'Strongly disagree' receives 5.

4. After the weight has been given to items, a total score for each subject is found by adding the weights earned by him on each item. Thus his total score is obtained after the weights are summated over all the statements. Since a subject's response to each item may be considered as his rating of own attitudes on a 5-point scale and his total score is obtained after all these weights are summated, the method is known as the method of summated ratings.

5. Finally, selection of items is done through the procedure of item analysis. Probably, this step of item analysis is the major step, which distinguishes it from Thurstone's method of equal-appearing intervals. As we have seen, Thurstone's method makes no use of item analysis in final selection of items. There are several methods of item analysis. Edwards (1957) has suggested the setting of two extreme groups—high and low—on the basis of the total score and finding out the significance of the difference between the means of two groups by the *t* test. The value of *t* will indicate the extent to which a given statement distinguishes between high and low groups. But other methods such as correlational methods may also be used in place of the *t* test.

In the method of summated ratings, it is customary to select 20 to 25 statements, which constitute the final attitude scale. As far as possible, half of the total statements should be favourable so that half of 'strongly agree' may receive the weight of 5 and the remaining half of total statements should be unfavourable so that half of 'strongly agree' may receive the weight of 1. This type of arrangement is necessitated to control certain response biases of subjects, which might be produced if only favourable statements or only unfavourable statements are included in the attitude scale.

The Likert method has also advantages and disadvantages. Its major advantages are mentioned below.

(i) Likert scales are easy to construct and it takes less time. This method is simpler and easier than the method of equal-appearing Intervals. Some empirical evidences are available to support this content. Rundquist & Sletto (1936, 5) have used Likert method in the construction of the attitude scale and expressed the belief that this method "... is less labourious than that developed by Thurstone." Edwards & Kenney (1946) have made a comparative study of Likert method and Thurstone method and have concluded that the time required in the construction of the attitude scale by equal-appearing Intervals is almost twice the time required by the method of summated ratings.

(ii) Scoring of Likert scale is easy as well. Statements on Likert scale are worded positively or negatively and subsequently, numerical weights are assigned to them. Subsequently, they are summed to yield total score. High total score indicates favourable attitude and low total score indicates unfavourable attitude.

(iii) Likert's summated ratings are the most common measurement format. The ease of application and simplicity of interpretation have increased the popularity of this measurement format in social science researches.

(iv) Likert's method of scaling possesses sufficient degree of flexibility. Here the investigator is free to include as many and as few items in his measure as he chooses. In this scaling, because each item is presumed to count equally in measuring the concerned phenomenon, increasing the number of statements increases the ability of the scale to reveal differences in the phenomenon measured.

Despite these advantages, some disadvantages or weaknesses have been reported in Likert's method of scaling as mentioned below.

(i) In Likert's method of summating ratings, it is assumed that each item or statement has identical weight in relation to every other item or statement. This is not necessarily a valid assumption. In fact, the different individuals may have a given attitude to the same degree, yet they may respond differently to different statements or items of the scale. Therefore, it is difficult, if not impossible, to ensure that each statement counts the same as every other item.

(ii) The validity of Likert's scaling is questionable. As we know, the process of deducing items from an abstract universe of traits is a logical one. Therefore, there always exists the possibility that some items may be wrongly included in the scale at any given time. The problem is to know that we are measuring exactly what we claim to measure.

(iii) In summated ratings, the persons receiving the same score on a measure don't necessarily possess the trait to the same degree. This obviously means that this method is never as precise as it claims to be. Its raw scores may be regarded as crude estimates at best.

Despite these disadvantages the method of summated ratings, as devised by Likert, has been successfully used in assessment of attitudes. Recently, this method has also been fruitfully used in assessing socio-economic status, intelligence, interest and special skills (Black & Champion 1976).

Guttman's Scale, or Cumulative Scale

Guttman's method of scale analysis or scalogram analysis differs considerably from the two methods of attitude scale construction discussed previously. The Guttman scale is based upon the methods of cumulative scaling and has been defined by Guttman (1950) as follows:

"We shall call a set of items of common content a scale if a person with a higher rank than another person is just as high or higher on every item than the other person."

Thus, if a set of statements with common content defines the Guttman scale, a person with higher score or rank than another person on the same set of statements will rank consistently higher than him on each statement in the set. For example, the following items illustrate the perfect Guttman's scale:

(a) My height is more than 5′. (b) My height is more than 5′ 3″.

(c) My height is more than 5′ 8″. (d) My height is more than 6′.

A person who responds with 'Yes' to item (d), will also be responding in 'Yes' to items (a), (b) and (c). All the four items are measuring the same dimension, that is, height and constitute what Guttman (1944,1945) called 'unidimensional scale.' Similarly, if a set of attitude statements measures the same attitude, they are said to constitute a unidimensional scale or a Guttman scale. According to Guttman, one advantage of the unidimensional scale is that from the total score of the person one can *reproduce* the pattern of his responses to the set of statements. Suppose, for example, that in the above example, 'Yes' is given a weight of 1 and 'No' is given a weight 0, then knowing that a person has secured a total weight of 4, we can say that he has responded 'Yes' to items a, b, c and d. Likewise, if a person gets a total weight of 3, he has responded 'Yes' to item a, b and c and 'No' to item d. Such prediction regarding the perfect reproducibility is true in a perfect Guttman scale only. In case of attitude, statements showing such a perfect reproducibility are rarely achieved because some degree of irrelevancy is always present.

The major steps in the Guttman scale may be enumerated as shown below.

1. A large number of statements are collected regarding the object of attitude. All statements seem to indicate the same attitude. This constitutes what Guttman calls a *universe of items.*

2. Out of these collected statements, a small number of items are selected. Usually, the number of selected items does not exceed 20. According to Guttman, the selection of a small number of items from the large number of possible items is dependent upon the *intuition* and *experience* of the investigator. These selected items must be of *homogeneous content.* Thus one should look for items in the Guttman scale which are, to a greater extent, the rephrasings of the same content. Guttman believed that item analysis was not an essential part of scale analysis for selection of items as we find in case of the Likert's scale.

3. Each statement may have two alternatives such as 'Agree–Disagree' or more than two alternatives such as 'Agree, Neutral and Disagree.' All these items are administered to a group of 100 persons who respond to each item.

4. All items are scored or weighted and a total score by adding the weights on all items is determined for each person.

5. On the basis of the total score, each subject is ranked from highest to lowest and is listed in a column. Each row indicates the responses of a subject to different items (see Table 13.20).

6. Subsequently, it is determined whether or not the responses to each item are in close agreement. In other words, those marking the response category (such as 'Agree'), which strongly indicates the quality being measured should show consistency with the higher total scores and those marking the response category (such as 'Disagree' or 'Neutral'), which poorly indicate the quality being measured should show consistency with the lower total scores. If this is reality, the scale is to be a homogeneous one and in this case from the person total score (or rank) alone, we can reproduce his response to each item. In a perfectly homogeneous test, the index of reproducibility will also be perfect and therefore, the coefficient of reproducibility will be one.

A case of perfect reproducibility has been demonstrated in Table 13.20 where responses of 10 subjects towards five items have been displayed. Each item has two response categories—'Agree' and 'Disagree'. The response category 'Agree' is scored with +1 and the other response category 'Disagree' is scored with 0. Subsequently, an attempt is made to evaluate the scalability of the items and for this purpose, there are several procedures.

We shall here describe the most common method adopted by Guttman (1947). The method is called the 'Cornell technique'. The question Guttman proposes to answer with this technique is whether the set of statements is unidimensional or multidimensional. Following this technique, a table is constructed in a way that for each response category of each statement one column is constructed and for each subject one row is constructed (see Table 13.20). Starting with the highest total score, we place a cross mark in the appropriate cell of the table for indicating responses given by each subject towards each item. Thus the table makes available all the

Table 13.20 Perfect Guttman scale of 5 statements responded by 10 subjects. +1 indicates favourable judgement and 0 unfavourable judgement

Subjects	Statements										Total Score
	1		2		3		4		5		
	+1 A	0 D	+1 A	0 D	+1 A	0 D	+1 A	0 D	+1 A	0 D	
1	×		×		×		×		×		5
2	×		×		×		×		×		5
3	×		×		×		×		×		5
4		×	×		×		×		×		4
5		×		×	×		×		×		3
6		×		×		×	×		×		2
7		×		×		×	×		×		2
8		×		×		×		×	×		1
9		×		×		×		×		×	0
10		×		×		×		×		×	0

responses of all subjects along with the total scores. Knowing the total score of a subject, we can reproduce his responses to each of the items. An inspection of Table 13.20 indicates that subject 1 has agreed to all the five statements. Therefore, his score is five. Subjects 2 and 3 have similarly responded. Subject 4 has responded `disagree' to item number one but has agreed with other four items. Subject 5 has responded `disagree' to item one and item two and has agreed to the remaining three items. Subject 6 has responded `disagree' to the first three items and has agreed to the last two items. Subject 7 has responded in a similar manner. Subject 8 has responded `disagree' to the first four items and has agreed with the last item, that is, the fifth item. Hence his total score is one. Subject 9 has earned the score of zero which means he has disagreed with all the five items. When the items form a perfect scale and vary in graduated fashion in intensity as items 1, 2, 3, 4, and 5 in Table 13.20, Guttman has named the resulting scale a perfectly reproducible scale which obviously means that a person's response pattern to all items in a set can be perfectly reproduced simply by knowing his total score on the scale. Such perfect reproducibility of scale can't be achieved by either the Thurstone method or the Likert method and hence, this may be considered as one advantage over these two scales.

In practice the ideal model of perfect scale is rarely shown, and therefore, it becomes necessary to examine the degree of reproducibility for the set of given attitude statements after making allowance for the errors. For this purpose, Guttman suggests that a *cutting point* should be located in the table. A cutting point may be defined as that point where the most common response is shifted from one category to another in the statement. Locating a cutting point sometimes creates difficulties. To ward off these difficulties, Guttman (1947) has given two points as guidelines. First, the cutting point for each statement should be located at a place where the error is minimised. Second, the cutting point should be located in a way that "no category should have more error in it than non-error." In case of a perfect scale or perfect reproducibility, the responses above the cutting point fall in one category and the responses below the cutting point fall in another category. This is clear from data presented in the Table 13.20. The cutting points have been shown in Table 13.20 with horizontal lines. Now we are presenting another set of data in Table 13.21 for the purpose of illustrating a nonperfect scale. The *cutting points* for each statement have been shown by horizontal lines in the Table 13.21, which displays the score matrix for 6 statements, each having two alternatives and answered by 10 subjects. In this table, the cutting points have been located for each statement according to the two-point suggestion of Guttman. The sum of each row of the score matrix yields the total score for each subject and these have been recorded in the last column. The first row at the bottom of Table 13.21 shows the frequency with which two categories of each statement have been responded. The sums for each column of the response category of 1 have been divided by the total number of subjects responding to this response category to obtain proportions p. The proportions of the subjects giving 0 responses will thus be equal to $1 - p = q$. The values of p and q have been shown in the second row at the bottom of the score matrix. Finally, errors for each response category have been counted and written at the bottom row of the score matrix. Here, errors indicate response inconsistencies. For the response category of 1 of the first statement, no response falls below the cutting points and hence, the error is zero, whereas for the response category of 0 one response falls above the cutting point but it should theoretically fall below the cutting point and hence, one error has occurred. Therefore, 1 has been written in the row of error against this category column. For the response category of 1 of the second statement, one response falls below the cutting point but it should theoretically fall above the cutting point and hence, one error has occurred. No response falls above the cutting point in the response category of 0 of the second statement and hence, no error has occurred. Similarly, error has been counted for each statement. The error for each statement has been added together which is equal to 8. We have a total of 60 responses (that is 10 × 6). The proportion of error is, therefore, equal to 8/60 = 0.133. Subtracting this value from unity, that is 1, we get the coefficient of reproducibility, which indicates the per cent accuracy with which the responses to various statements can be reproduced. Thus, coefficient of reproducibility $= 1 - 0.133 = 0.867 = 0.87$. Thus, the equation for coefficient of reproducibility becomes:

$$\text{Coefficient of reproducibility} = 1 - \frac{\text{number of errors}}{\text{number of responses}} \qquad (13.27)$$

Table 13.21 An illustration of the Cornell technique in the Guttman scale

	Statements												
	1		2		3		4		5		6		Total
Subjects	1	0	1	0	1	0	1	0	1	0	1	0	Scores
1	×		×		×		×		×		×		6
2	×		×		×		×		×		×		6
3	×		×		×		×		×		×		6
4	×		×		×			×	×			×	4
5	×			×	×			×		×	×		3
6		×		×	×			×	×		×		2
7	×			×		×	×			×		×	2
8		×	×			×		×	×		×		2
9		×		×	×			×		×		×	1
10		×		×		×		×		×		×	0
f	6	4	5	5	7	3	4	6	6	4	6	4	
p and *q*	0.6	0.4	0.5	0.5	0.7	0.3	0.4	0.6	0.6	0.4	0.6	0.4	
error	0	1	1	0	1	0	1	0	2	0	1	1	$\Sigma = 8$

According to Guttman, if the coefficient of reproducibility is below 0.90, no cumulative scale is said to exist or if it is between 0.85 to 0.90, a 'quasi scale' is said to exist. Thus, for Guttman, the coefficient of reproducibility must be at least 0.90 for constituting the cumulative scale.

Guttman scale has some advantages and disadvantages. Its major *advantages* are as follows:

(i) The Guttman scale clearly demonstrates the unidimensionality of items. Such unidimensionality is assessed neither by the Likert scale nor by the Thurstone scale.

(ii) The unidimensionality and scalability of the scale enable the researchers to identify any inconsistencies in the responses and probable untruthful replies given by the subjects.

(iii) In the Guttman scale, the person's response pattern can be easily reproduced with a knowledge of his total score on the scale. This type of advantages is not found in case of the Thurstone scale or the Likert scale.

(iv) Researchers have shown that the Guttman technique of scaling is relatively easy to use when the number of dichotomous items (such as agree–disagree) is 12 or less than that. When the number of items exceeds 12, the technique becomes cumbersome.

The Guttman scale has, however, some *disadvantages*. The major ones are given below.

(i) The Guttman scale is not well suited to those items which have three or more than three response categories. Although some psychologists have applied this scaling technique in such a situation, it is very cumbersome and tedious to proceed with this technique.

(ii) The Guttman scale can't be used appropriately in situations where the number of items exceeds 12 and the number of subjects is more than 100. In such scalogram analysis, scoring and error determination will prove to be a Herculean task.

(iii) The Guttman scaling technique does not provide as extensive continuum for scaling as we find in the case of Likert's and Thurstone's scaling technique.

Despite these limitations, the Guttman scale has been successfully used in attitude assessment. Besides, the technique has also been used in opinion studies of a political, economic and social nature. It may also be used in combination with Likert's and Thurstone's scale to encompass a wide variety of attitude assessment.

Relationship between Attitude Scale and Measurement Scale

There is a defined relationship between different attitude scales and measurement scales. It is very important to know the link or relationship between attitude scales and measurement scales because this will help in interpreting the scores on attitude scales. This relationship may be explained as under:

(i) Likert scale is based on ordinal scale.
(ii) Thurstone scale is based on interval scale.
(iii) Guttman scale is based on ratio scale.

☑ Review Questions

1. Describe and compare the scaling techniques of rank order and paired comparison.
2. Critically evaluate the advantages and disadvantages of the equal-appearing intervals method.
3. Discuss the general procedure involved in a scale construction.
4. Discuss the contributions of psychophysical methods to the development of psychological research.
5. Describe the steps involved in developing a Likert type of scale to measure attitude towards family planning.
6. Explain the procedure of scaling by the method of paired comparison.
7. Discuss (i) Method of constant stimuli, and (ii) Guttman scale.
8. What do you mean by a unidimensional scale? Outline the steps involved in a unidimensional scale as suggested by Guttman.
9. Describe with examples, the major steps involved in construction of a scale by method of equal-appearing intervals.
10. Compare and contrast the method of summated ratings and the method of equal-appearing intervals. Which of the two do you consider better for developing an opinionnaire? Give reasons.
11. Write short notes on:
    (a) RL and DL
    (b) Guttman–Cornell technique
    (c) Psychological methods
    (d) Steven's power law

✣

Part Three

PRINCIPLES OF RESEARCH METHODOLOGY

14

SAMPLING

CHAPTER PREVIEW

- Meaning and Types of Sampling
  - (A) Probability Sampling Methods
  - (B) Nonprobability Sampling Methods
  - (C) Mixed Sampling
- Need for Sampling
- Fundamentals of Sampling
- Principles of Sampling
- Aims of Sampling
- Factors Influencing Decision to Sample
- How Large Should a Sample Be?
- Methods of Drawing Random Samples
- Simple Random Sample
- Stratified Random Sample
  - (A) Proportionate Stratified Random Sampling
  - (B) Disproportionate Stratified Random Sampling
- Area (or Cluster) Sampling
- Quota Sampling
- Purposive or Judgemental Sampling
- Accidental Sampling
- Snowball Sampling
- Saturation Sampling and Dense Sampling
- Double Sampling
- Mixed Sampling
- Requisites of a Good Sampling Method
- Common Advantages of Sampling Methods
- Sampling Distribution
- Sampling Error

MEANING AND TYPES OF SAMPLING

Nearly all researches—experimental and nonexperimental—in the behavioural sciences, particularly in the fields of psychology, sociology and education—draw some inferences regarding a well-specified and identifiable group on the basis of some selected measures. The well-specified and identifiable group is known as a *population* or *universe* and the selected number of persons or objects is known as a *sample*. The generalized conclusions are technically known as the *statistical inferences*. A population, therefore, may be defined as any identifiable and well-specified group of individuals. All primary school teachers, all college teachers, all university students, all housewives, etc., are examples of populations. A population may be *finite* or *infinite*—a finite population is one where all the members can be easily counted; an infinite population is one whose size is unlimited and therefore, its members cannot be counted. The population of university teachers is an example of a finite population and the population of fishes in a river is an example of an infinite population because the former can be counted whereas the latter cannot be counted.

Likewise, a population may be *real* or *imaginative*—a real population is one which actually exists and an imaginative population is one which exists only in the imagination. In psychological and educational research, on many occasions, the population is imaginative. A measure based upon the entire population is called a *parameter.*

A sample is any number of persons selected to represent the population according to some rule or plan. Thus, a sample is a smaller representation of the population. A measure based upon a sample is known as a *statistic*.

Before we define the different types of samples (or the methods of sampling), it is essential to define the term 'probability', which is the base of sampling theory. The general meaning of probability is *less than certain* and for which there exists some evidence. In sampling theory, the term 'probability' is used as equivalent to the relative frequency. Thus when we say that the probability of a tail on a single toss of a coin is 1/2, it is meant that when we make several tosses, the relative frequency of a tail will be about 1/2 or 0.5. If one says that the probability of having of a male child is 0.8, it is meant that on previous occasions the relative frequency of the birth of a male child has been 0.8. Probability may be expressed in terms of a fraction or in decimal numbers.

Following Blalock (1960), most sampling methods can be categorized into two—

(A) Probability Sampling Methods

(B) Nonprobability Sampling Methods

A discussion of these two is given below.

(A) Probability Sampling Methods

Probability sampling methods are those that clearly specify the probability or likelihood of inclusion of each element or individual in the sample. Technically, the probability sampling methods must satisfy the conditions given below.

(i) The size of the parent population or universe from which the sample is to be taken, must be known to the investigator.

(ii) Each element or individual in the population must have an equal chance of being included in a subsequent sample.

(iii) The desired sample size must be clearly specified.

If, for example, a researcher knows that the population which he is going to study contains 500 elements, or individuals, and if he knows that all the elements (or individuals) are accessible and may be included in a subsequent sample, it can be said that each element (or individual) in the population has an equal chance, that is, 1/500 of a chance of being selected. This constitutes the probability sampling method. In practice, however, sometimes researchers are not able to know for certainty that conditions (i) and (ii) will be satisfied. Sometimes the population studied is so large as to be considered infinite and unknowable for all important and practical purposes.

The positive point of the probability sampling method is that the obtained samples are considered representative, and hence, the conclusions reached from such samples are worth generalization and are comparable to similar populations to which they belong.

The negative point of the probability sampling method is that a certain amount of sampling error exists because the researcher has only a limited element of the entire population. Sampling error refers to the degree to which the sample characteristics approximate the characteristics of the parent population. The *smaller* the sample, the *greater* the sampling error.

The major probability sampling methods are the following:

1. Simple random sampling
2. Stratified random sampling
    - (a) Proportionate stratified random sampling
    - (b) Disproportionate stratified random sampling
3. Area or cluster sampling

(B) Nonprobability Sampling Methods

Nonprobability sampling is one method in which there is no way of assessing the probability of the element or group of elements of the population being included in the sample. In other words, nonprobability sampling methods are those that provide no basis for estimating how closely the characteristics of a sample approximate the parameters of the population from which the sample had been obtained. This is because nonprobability samples don't use the techniques of random sampling. Important techniques of nonprobability sampling methods are:

1. Quota sampling
2. Accidental sampling
3. Judgemental or purposive sampling
4. Snowball sampling
5. Saturation sampling
6. Dense sampling

(C) Mixed Sampling

- Systematic sampling

Figure 14.1 presents the details of classification of sampling.

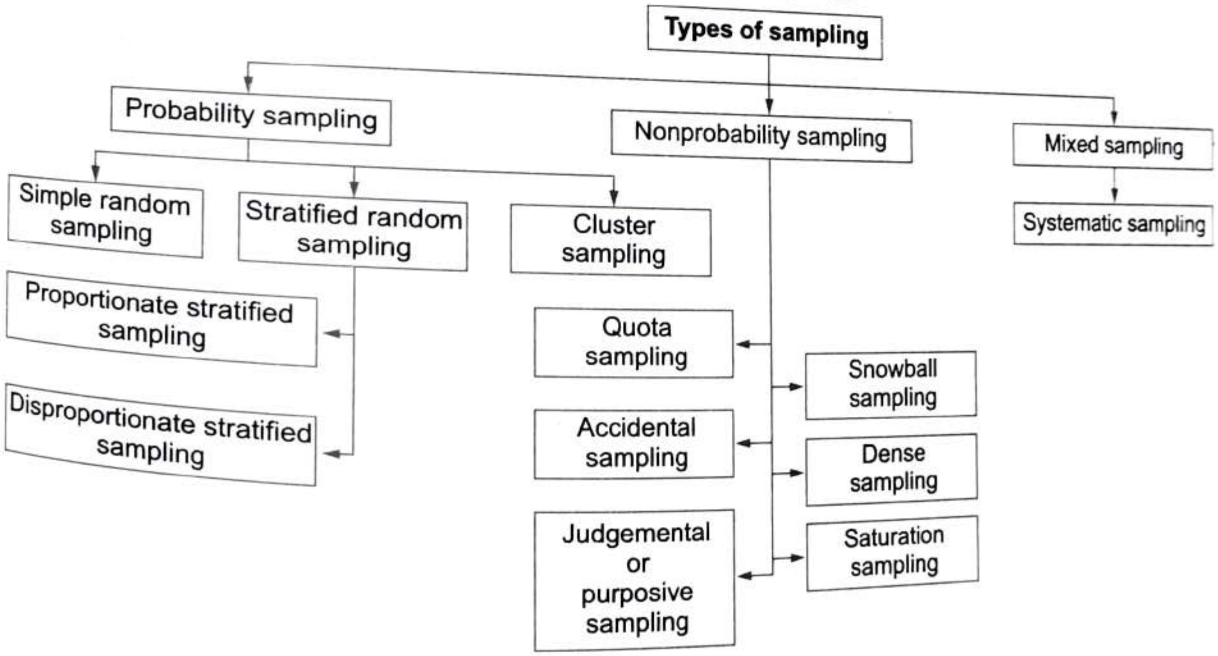
Types of sampling
Probability sampling
Nonprobability sampling
Mixed sampling
Simple random sampling
Stratified random sampling
Cluster sampling
Systematic sampling
Proportionate stratified sampling
Disproportionate stratified sampling
Quota sampling
Accidental sampling
Judgemental or purposive sampling
Snowball sampling
Dense sampling
Saturation sampling

Fig. 14.1: Details of types of sampling

NEED FOR SAMPLING

Sampling is needed for a variety of reasons. Some of the important reasons are mentioned below.

(i) Sampling saves time as well as money. A research study based on sampling is completed within a lesser time and incurs less expenditure than a study based upon other criteria.

(ii) A research study based upon sampling is generally conducted by trained and experienced investigators. As such, it provides accuracy in measurement and testing.

(iii) Sampling is also needed because it enables the researcher to estimate the sampling error and that way, it helps in getting information regarding some characteristics of the population.

(iv) Sampling is needed because it remains the only way when a population/universe contains infinitely many members.

(v) Sampling helps in making correct and scientific judgement about the population for which generalization is to be made after the completion of the study.

FUNDAMENTALS OF SAMPLING

There are some fundamental concepts related to sampling that need to be explained. Such important fundamental concepts are *universe/population*, *sampling frame*, *sampling design*, *sampling error*, *sampling distribution*, *statistic(s)* and *parameter(s)*, *confidence interval* and *significance level*. The meaning of universe/population, sampling error and sampling distribution has been explained elsewhere in this chapter. Consequently, we shall discuss the meaning of only the remaining fundamentals.

(1) Statistic(s) and Parameter(s)

A statistic is a numerical value based upon sample, whereas parameter is the numerical value based upon population. For example, when we calculate mean from a sample, this is called statistic because it describes the characteristics of a sample. When the same mean is calculated from population, it is called parameter because it describes the characteristics of population. Any sampling analysis aims to obtain the estimate of a parameter from a statistic.

(2) Sampling frame

The elementary units form the basis of the sampling process and such units are called sampling units. A list containing all such sampling units is called a *sampling frame*. In other words, it can be said that the sampling frame consists of a list of items from which sampling is to be finally drawn. In some cases, it becomes impossible to draw a sample directly from a population. As such, this frame is constructed by the researchers for the purpose of the study or it may consist of some existing list of population. For example, the researcher can use a telephone directory as a frame for conducting an opinion survey in a particular town.

(3) Confidence interval and Significance level

The confidence interval is the expected percentage of times stipulating that the actual value will fall within the stated precision limits. For example, if we take a confidence interval of 95%, then obviously, it means that there are 95 chances in 100 that the sample results present the true characteristics of the population within a specified precision range against 5 chances in 100 that it does not. Thus confidence interval indicates the likehood that the answer will fall within the precision range. Significance level, on the other hand, is the level that indicates the likehood that the answer will fall outside that range. If the confidence level is 95% then the significance level is 100 – 95 = 5% or .05 and if the confidence level is 99%, the significance level is 100 – 99, that is 1%. It would not be out of place to mention here that the area of the normal curve within precision limits for the specified confidence level constitutes the acceptance region, and the area of curve outside these limits in either direction constitutes the rejection regions.

(4) Sampling design

By a sampling design, we mean a plan for obtaining a sample from the sampling frame. In other words, it refers to the procedure the researcher would adopt in selecting some sampling units from which inferences about the population is drawn. A good sampling design must possess the following characteristics:

(i) The sample design must provide a sample which would be the true representative of the population.
(ii) The sample design should be such that it may result in small sampling error.
(iii) The sample design must allow for controlling systematic bias in the sampling.
(iv) The sample design must be usable in light of the funds available for the research study.
(v) The sample design must be such that the results of the sample study can be applied in general, for the population with a reasonable confidence level.

There are two broad types of sample designs: *probability sampling* and *nonprobability sampling*. These will be discussed in detail later on in this chapter.

PRINCIPLES OF SAMPLING

There are three principles of sampling that guide the theory of sampling. These principles are as under:

1. *In most cases of sampling, there will be difference between sample statistics and the true population parameter and this is attributable to the selection of units in the sample.*

To explain this principle in a better way, let us take an example. Suppose there are five individuals and you know their age.

Individuals	**Age (in years)**
A	20
B	22
C	19
D	20
E	24
	Total = 105

$$\text{Mean age} = \frac{105}{5} = 21 \text{ years}$$

Now let us suppose further that you want to select a sample of two persons for making an estimate of the mean age of these five persons (population mean or parameter mean). If you adopt the theory of probability, there can be 10 possible combinations of two: AB, AC, AD, AE, BC, BD, BE, CD, CE and DE. Now let us calculate the average (or mean) age of each of these pairs.

	Difference between sample mean and population mean
1. $A + B = 20 + 22 = 42/2 = 21$	0
2. $A + C = 20 + 19 = 39/2 = 19.5$	–1.5
3. $A + D = 20 + 20 = 40/2 = 20$	–1
4. $A + E = 20 + 24 = 44/2 = 22$	1
5. $B + C = 22 + 19 = 41/2 = 20.5$	–0.5
6. $B + D = 22 + 20 = 42/2 = 21$	0
7. $B + E = 22 + 24 = 46/2 = 23$	2
8. $C + D = 19 + 20 = 39/2 = 19.5$	–1.5
9. $C + E = 19 + 24 = 43/2 = 21.5$	0.5
10. $D + E = 20 + 24 = 44/2 = 22$	1

Now look at the average ages calculated on the basis of the sample of two individuals (sample statistics). If you make comparison of these averages with mean or average of the all five individuals (population mean, that is, 21 years), you will find that only in case of two combinations, that is, AB and BD, there is *no* difference between the mean of the sample and mean of the population. In case of remaining eight combinations, there is difference. This difference is known as *sampling error* whose size varies markedly. This analysis well illustrates the first principle of sampling stated above.

2. *The second principle is that the greater the size of sample, the more accurate will be the estimate of the population value or parameter.*

This principle of sampling obviously relates the size of the sample with the accuracy of estimate of the population mean. To illustrate this principle, let us carry the above example further. Now let us take a bigger size of sample, that is, of three individuals in place of two individuals and see what happens. There can be seven possible combinations of three individuals, out of a population of five individuals.

	Difference between sample mean and population mean
1. $A+B+C=20+22+19=61/3=20.3$	–0.7
2. $A+B+D=20+22+20=62/3=20.6$	–0.4
3. $A+B+E=20+22+24=66/3=22$	1
4. $B+C+D=22+19+20=61/3=20.3$	–0.7
5. $B+D+E=22+20+24=66/3=22$	1
6. $B+C+E=22+19+24=65/3=21.6$	0.6
7. $C+D+E=19+20+24=63/3=21$	0

Now, make a comparison of the differences when the sample size was of two individuals to the differences when the sample size is of three individuals. Difference between sample mean and population mean in case of sample size of two individual lies in between 0.5 to 2.00 whereas in case of the sample size of three, this difference is reduced and lies between –0.4 to –0.7. Thus, it is clear that with enhancement in sample size, the sampling error is reduced and accuracy of estimate to population mean is enhanced.

3. *The third principle is that the greater the difference in the variable under study in a population for a given sample size, the greater will be the sampling error, that is, greater will be the difference between sample satistics and true population parameter.*

This principle clearly states that when the population is heterogeneous or markedly varies with respect to the variable under study, there will be the difference between sample statistics and population true value. Let us take an example to illustrate this. Suppose we know that of the five individuals, the age of A is 20 years, that of B is 30 years, that of C is 25 years, that of D is 40 years and that of E is 45 years. Then the average age of these five individuals who markedly differ in terms of age will be

Individuals	**Age (in years)**
A	20 years
B	30 years
C	25 years
D	40 years
E	45 years
	Total = 160 years

$$\text{Mean age} = \frac{160}{5} = 32 \text{ years}$$

If we take a sample of two individuals, their mean age and differences from the true population mean (32 years) will be as under:

	Differences between sample mean and population mean
1. $A + B = 20 + 30 = 50/2 = 25$	−7
2. $A + C = 20 + 25 = 45/2 = 22.5$	−9.5
3. $A + D = 20 + 40 = 60/2 = 30$	−2
4. $A + E = 20 + 45 = 65/2 = 32.5$	0.5
5. $B + C = 30 + 25 = 55/2 = 27.5$	−4.5
6. $B + D = 30 + 40 = 70/2 = 35$	3
7. $B + E = 30 + 45 = 75/2 = 37.5$	5.5
8. $C + D = 25 + 40 = 65/2 = 32.5$	0.5
9. $C + E = 25 + 45 = 70/2 = 35$	3
10. $D + E = 40 + 45 = 85/2 = 42.5$	10.5

If you look at the difference, it varies from 0.5 to 10.5, thus producing a bigger sampling error. This happens because of the fact that the individuals in the population differ markedly in terms of variable under study, that is, age. When they were more or less of the same age, as we find in the example of the first principle, the range of difference was in between 0.5 to 2.00 only, thus showing a lower degree of sampling error.

AIMS OF SAMPLING

There are certain aims in selecting a sample. These aims are:

(i) To achieve maximum precision in estimating about population values within a given sample size

(ii) To avoid bias in selecting a sample

The common observation is that bias in selecting a sample occurs in the following situations:

(a) When sampling is done by nonrandom method, that is, when sampling is influenced by human choice, the bias occurs.

(b) When a section of population refused to cooperate or somehow is not available

(c) When the sampling frame does not cover the whole population

FACTORS INFLUENCING DECISION TO SAMPLE

Any behavioural researcher takes some decision regarding the sampling plan at the time of formulating the research design. His decision to sample is influenced by at least three factors:

1. Size of the population
2. Cost involved in obtaining the elements
3. Convenience and accessibility of the elements

These three may be discussed as follows:

1. Size of the Population

Decision to sample is directly influenced by the size of the population. If the population is small, say, for example, it consists of 300 individuals, the investigator may decide to include all in his study and, therefore, sampling may not be done. On the other hand, if the size of the population is large, say, it consists of 10,000 individuals, he may decide to select a limited number of individuals from the population of 10,000 individuals. Obviously, then it can be said that as the population of elements (or individuals) becomes larger, sampling becomes increasingly very important. It may also be mentioned here that the size of the population is a relative matter. What one investigator regards as a large population, the other may regard it as a small one. In fact, no clear-cut guidance exists for making distinction between large population and small population.

2. *Cost Involved in Obtaining the Elements*

The investigator is also influenced by the cost likely to be incurred in obtaining the elements from the population. If sampling involves a bigger cost which the investigator can't meet, the decision to sample may be postponed. On the other hand, if sampling involves a cost which the investigator can readily meet, the sampling work is greatly facilitated.

3. *Convenience and Accessibility of the Elements*

The decision to sample is also influenced by the convenience and accessibility of the elements. Sometimes the investigator may have to deal with a problem with respect to which the elements may not be conveniently available. For example, if the problem deals with investigating the causes of premarital sex relationship among college girls, the investigator may not find it convenient to trace out such girls who could readily tell about their affairs openly. In such cases, sampling may be faulty. On the other hand, some investigators may have access to facilities and staff where a large amount of data could be easily handled. These investigators dare sample even in those cases where the resulting data are complicated and complex.

Obviously, the decision to sample effectively is influenced by the size of the population, the anticipated cost of the study and the convenience and accessibility associated with the elements.

HOW LARGE SHOULD A SAMPLE BE?

It is often asked by student, "how large the size of sample be in my study?" The best answer is that it *depends* upon many factors. In fact, a good researcher's decision about the best sample size depends upon three factors.

(a) Degree of accuracy required

(b) Degree of variability in the population

(c) Number of different variables examined simultaneously in data analysis

Other things being equal, the researcher needs larger samples if he wants higher degree of accuracy, if the population has a great deal of heterogeneity or if he wants to examine many variables in the data analysis simultaneously. Smaller samples are considered good when less accuracy is acceptable, when the population is homogeneous or when the researcher wants a few variables only to be examined at a time. In general, the best principle of determining sample size is that if the population is small, a bigger sampling ratio has to be for an accurate sample. Larger sample permits smaller sampling ratios for equally good samples because as the population size grows, the returns in accuracy for sample size shrink. Let us take an example. For small populations (under 1,000), a researcher must have a larger sampling ratio (that is, about at least 30%). Thus, a sample size of about 300 is required for a high degree of accuracy. For moderately large populations (10,000), a sampling ratio of about 10 per cent is needed for achieving equal accuracy or a sample size of 1000 can be accurate. For larger populations (over 160,000) a smaller sampling ratio, that is, 1 per cent may yield equally accurate results or a sample of 1600 can be accurate. These are, however, approximate sizes and practical limitations such as cost, time, etc., also play a role in taking a decision by the researcher.

Till the second decade of the 20th century, statisticians believed that samples should be large so that the normal probability table could be easily used to estimate sampling error as explained by *Central limit theorem*. However, the research work of Sealy Gosset in 1915 clearly provided data on the probability distribution of small sample means (Student's t distribution) that led to the effective use of small samples. In general, samples of 30 are usually considered large samples and those fewer than 30, are considered small samples. At sample size of 30, the magnitude of student's t values for small samples come nearer to the Z critical values of the normal probability table for large samples. Samples that are larger than necessary may also produce problems like higher required cost as well as administrative and logistic problem.

Estimation, Confidence Intervals and Confidence Limits

One of the basic problems in inferential statistics is the estimation of parameters of population from statistics calculated from a sample. Traditionally, this has been very important in survey research but in recent years, it is also becoming important in experimental research and can even serve as an alternative approach to hypothesis testing. In fact, this problem involves two subproblems.

(a) Point estimation
(b) Interval estimation

When the researchers estimate parameters using single sample value, these estimates are known as *point estimates*. Thus in point estimation, a single sample value drawn from a population provides an estimate of the population parameter. The question is: how good such an estimate is? Suppose that if the population mean is known to be 80, would a sample mean of 60 constitute a good estimate? What about sample mean of 50, 75, 100, 105, etc.? Under what conditions do we consider an estimate good? Since the researcher knows that population parameters are virtually never known and that we generally use samples to estimate these parameters, is there any means to determine the amount of error the researcher is likely to make? A way to get at this question is to put another related question: how much do means from a distribution of means vary? One indication of such variation is to compute the standard deviation of the distribution of means. Like any standard deviation, how much scores (here sample means) in the distribution typically vary. Since the interest of the researcher is in making an estimate of the population mean, any variation can be easily taken as an error.

The researcher can also estimate the range of possible values that he confidently believes includes the population parameter. In other words, we can also estimate the range of possible means that are likely to include the population mean. This is known as *interval estimation*. Social scientists generally prefer interval estimates over point estimates. Let us illustrate the meaning of interval estimation with the help of day-to-day example. Suppose you are forced to guess the weight of a man based on physical inspection. Since you don't know the population value (the man's true weight), you are likely to base your impression on his close physical inspection. You may say that the man weighs 60 kg. If you are asked, "How confident you are that he weighs exactly 60 kg?", you would probably reply, "I doubt that he weighs exactly 60 kg. However, I feel reasonably confident that he weighs between 55 kg to 65 kg." In fact, in doing this, you have estimated the weight of the man in terms of interval between 55 kg to 65 kg. One important point to be noted here is that the greater the size of the interval, the greater is the feeling of certainty that true value is encompassed between these limits. The interval within which we consider the hypothesis tenable is called as *confidence interval* and limits defining the interval are referred to as *confidence limits*.

Now let us look at a sample problem and apply our statistical concepts to the interval estimation of confidence levels. Suppose mean of a sample is 120 with standard error of mean 6. Now in such a situation if the researcher randomly takes a mean from this distribution, there is a 34:1 probability that mean will be between 120 and 126 (one standard error above 120). This happens because the distribution of means is a normal curve, standard error is 1 standard deviation on that curve and we know that 34% of a normal curve is between mean and 1 standard deviation above the mean. Likewise, the researcher could also reason that another 34%

should be between 120 and 114 (1 standard error below 120). Putting these two together, the researcher has a range of 114 to 126 that he is 68% confident to include population mean if the sample was randomly taken from this population. We call it 68% confidence interval and the upper confidence limit here is 126 and the lower confidence limit is 114. Normally, however, the researcher wants to be more than 68% confident about the estimates. The standard practice is to use 95% or even 99% as confidence intervals. For the 95% confidence interval, the researcher wants an area in normal curve on each side between mean and z-score that includes 47.5%. The normal curve table clearly shows this to be at 1.96z. In terms of z-scores, the 95% confidence interval is from -1.96 to $+1.96$ on distribution of means. If we change these z-scores to raw scores in the above example, the lower limit would be $(-1.96)(6) + 120 = -11.76 + 120 = 108.24$ and the upper limit would be $(1.96)(6) + 120 = 11.76 + 120 = 131.76$. This shows that the true population mean for the given sample is between 108.24 and 131.76. Likewise, for 99% confidence interval the researcher uses the z- score for the middle 99% of the normal curve. This comes to be ± 2.56. Changing this to raw scores, the 99% confidence interval will be from 104.64 to 135.36.

METHODS OF DRAWING RANDOM SAMPLES

Randomness is one of the vital points in sampling, particularly, probability sampling (Peatman 1947). Randomness means that all elements in a given population have an equal and independent chance of being included in the sample. There are several methods through which the investigator maintains randomness in his sampling plans. Some of these methods are given below.

1. Fishbowl Draw Method
2. Using Table of Random Numbers
3. Method of Computer-determined Randomness

A discussion on these methods is given below.

1. Fishbowl Draw Method

This is a very simple method through which the elements from the population can be selected randomly. In this method the entire elements (or individuals) of the population are numbered on slips of paper of equal size, colour, etc. All these slips are folded in one and the same way and are put in a container or bowl. After mixing the slips thoroughly, the investigator, in a blindfolded position, selects one number at a time until the desired sample size is obtained.

Although this method is very simple for random selection of samples, it has some limitations.

(i) This method can't be applied where the size of the population is large. It would be a cumbersome and tedious task to number each element of a population of, say, fifteen thousand cases. Moreover, mixing them in the container or bowl will also pose a big problem.

(ii) This method of random selection is too simple a method to be considered as a scientific one. Here the investigator may select some numbers of slips from the bowl while purposefully excluding the others.

2. Using the Table of Random Numbers.

For random sampling the use of the table of random numbers is considered appropriate, easy and scientific. The table of random numbers consists of a continuous row-column sequence of numbers which don't appear in any particular sequence, nor does any number appear more frequently than the other one. For using the table of random numbers the researcher specifies the number of elements in the population and then, numbers them from 1 to *N* where *N* is the total number of elements in the population. Suppose the population size is of 500 and he intends to select 50 cases randomly from this population. He would enter the table at any point. He may move systematically to the right, left, up, down or diagonally, skipping the numbers that are too large and also those that have already been drawn. (This latter event does take place occasionally.) He would thus keep on moving systematically through the table until he has selected the random sample of 50 elements.

In Table 14.1, a random selection of 30 cases from a population serially numbered from 1 to 80 has been illustrated. Twelve numbers have been omitted. Numbers 85, 84, 97 and 95 have been omitted because they exceed 80 and numbers 03, 74 and 12 have been omitted because they have duplicated the previous selections. Number 00 has been omitted because our population number starts from 01.

Table 14.1 Table of random numbers and its illustrative use

64755	83885	84122	25920	17696	
10302	52289	77436	34430	38112	
71017	98495	51308	50374	66591	
60012	55605	88410	34879	79655	
Sample (*N* = 30 has been taken from a serially numbered population of 80)					
64	59	28	12	85	55
75	20	97	71	03	60
58	17	74	01	74	
38	69	36	79	66	
85	61	34	84	59	
84	03	43	95	16	
12	02	03	51	00	
22	52	81	30	12	

The advantage of using the table of random numbers is that it is easily accessible to the researchers and requires no formal training in using it. However, the disadvantage is that it can't be easily and constructively used when the size of the population exceeds 5 digits.

3. *Method of Computer-determined Randomness.*

This method is used when the size of the population is large. The data is fed in the computer to obtain a random number of elements corresponding to the elements in the population. The investigators note that this method is easier to adopt when computer facilities are available.

Of these three methods, the use of the table of random numbers is more popular for ensuring randomness in the method of selecting a sample.

SIMPLE RANDOM SAMPLE

A simple random sample (also known as an *unrestricted random sample*) may be defined as one in which each and every individual of the population has an equal chance of being included in the sample, and also the selection of one individual is in no way dependent upon the selection of another individual. For example, if we are to select a sample of 10 students from the seventh grade consisting of 40 students, we can write the name (or roll number) of each of the 40 students on separate slips of paper—all equal in size and colour—and fold them in a similar way. Subsequently, they may be placed in a box and reshuffled thoroughly. A blindfolded person, then, may be asked to pick up one slip. Here, the probability of each slip being selected is 1/40. Suppose that after selecting the slip and noting the name written on the slip, he again returns it to the box. In this case, the probability of the second slip being selected is again 1/40. But if he does not return the first slip to the box, the probability of the second slip becomes 1/39. When an element of the population is returned to the population after being selected, it is called *sampling with replacement* and when it is not returned, it is called *sampling without replacement*. Sampling with replacement is wholly feasible except in certain situations where it is seldom used (Cochran, 1963, 19). If sampling with replacement is used, the chance of the same case being selected more than once is increased. In such a situation, the repeated cases may be ignored as is done if a table of random numbers is used in making a selection of cases. Thus random sampling may be defined as one in which all possible combinations of samples of fixed size have an equal probability of being selected.

The major difference between *sampling with replacement* and *sampling without replacement* is mainly concerned with the number of possible samples of size *n* that could be theoretically drawn. In the case of sampling with replacement, the number of possible samples of size *n* would be greater than the number of possible samples of the same size (from same population) in case of sampling without replacement. This difference can be illustrated through an example. Suppose the size of the population consists of 4 persons, who are named as A, B, C and D. Further, suppose that the investigator wants to select samples of size 2 through the procedure of sampling without replacement. In such a situation, the investigator can maximally draw six samples of size 2 from the population of 4. This could be accomplished with the help of the following equation:

$$\binom{N}{n} = \frac{N!}{(N-n)!\,n!} \tag{14.1}$$

where, N = the size of parent population
n = the size of the sample
$!$ = factorial

In the above example where $N = 4$ and $n = 2$, the maximum number of sample size of 2 would be 6 as under:

$$\genfrac{}{}{0pt}{}{4}{2} = \frac{4!}{(4-2)!\,2!} = \frac{4\times3\times2\times1}{2\times1\times2\times1} = \frac{24}{4} = 6$$

Similarly, where $N = 5$, we can have 10 samples of size 2 as shown:

$$\genfrac{}{}{0pt}{}{5}{2} = \frac{5!}{(5-2)!\,2!} = \frac{5\times4\times3\times2\times1}{3\times2\times1\times2\times1} = \frac{120}{12} = 10$$

But from the same population, we can have 5 samples of size 4 as under:

$$\genfrac{}{}{0pt}{}{5}{4} = \frac{5!}{(5-4)!\,4!} = \frac{5\times4\times3\times2\times1}{1\times4\times3\times2\times1} = \frac{120}{24} = 5$$

But if the investigator has decided to proceed with the technique of sampling with replacement, he can derive the likely number of samples from the given population with the help of the following equation:

$$N^{n} \tag{14.2}$$

(where meaning of N and n is identical to the meaning in the case of Equation 14.1).

Suppose the size of the population is 4 and the size of the sample is 2. In such a situation the investigator, following the technique of sampling with replacement, can maximally draw 16 samples, that is, $4^2 = 4 \times 4 = 16$. If the four members of population are named as A, B, C, and D, the sixteen samples of size 2 would be

AA AB AC AD,
BA BB BC BD,
CA CB CC CD,
DA DB DC DD

The case of AA, BB, CC and DD combinations reflects the fact that in sampling with replacement, an element or individual once drawn can be drawn again. In actual practice, such cases are ignored.

There are some advantages and disadvantages of simple random sampling. The major *advantages* of simple random sampling are given ahead.

1. A sample prepared on the basis of simple random sampling plan is regarded as the representative of the population from which it was drawn. This is because in such a sampling plan, all the elements in the population have an equal and independent chance of being included in the sample. A sample drawn in such a way that does not ensure equal chance for all elements to be included in the population, rather increases or decreases the likelihood of an element being included, the resultant sample is called a *biased sample.*

2. In simple random sampling, the investigator need not know the true composition of the population beforehand. Such a sample theoretically reflects all important characteristics and segments of the population.

3. With a view to understanding and application, simple random sampling is the easiest and simplest technique of all probability sampling plans.

4. Simple random sampling serves as a foundation upon which all other types of random sampling are based because this method of sampling can be readily applied in conjunction with all other probability sampling plans.

5. In simple random sampling, the sampling error associated with any given sample drawn can easily be assessed.

6. In simple random sampling, the investigator does not commit classification errors because he need not know thoroughly the population characteristics prior to selection of the sample. Classification error means the error which results from the improper classification of population characteristics or segments.

However, simple random sampling has also some *disadvantages* as given below.

1. One of the major disadvantages is that simple random sampling does not ensure that those elements which exist in small numbers in the population will be included in the given sample. This is a very serious limitation. Suppose, in a population of 1,000 people, only 12 persons possess the X trait and the investigator is to draw only 50 cases from 1,000 people. If the investigator wants to include some persons possessing the trait X in his sample, chances are very slim that such persons would be drawn.

2. Another disadvantage of simple random sampling is that it does not fully exploit the knowledge the investigator has concerning the segments of the population. Suppose, for example, the investigator knows the proportion of males and females in the population, proportion of graduates and undergraduates, proportion of persons belonging to high SES and low SES persons in population, etc., he can't utilize these items of information while drawing a sample. This mars the quality of representativeness of the sample being drawn.

3. In the case of simple random sampling, the sampling error of a sample of size *n* is greater as compared with the sampling error incurred in the case of a stratified random sample of the same size. This is because the heterogeneity in the case of a random sample is greater than the same in the case of a stratified random sample. In stratified random sampling, the sample drawn becomes somewhat typical of the population because it is more or less proportionate to some

known characteristics of the population. This fact is ignored in simple random sampling. Hence, the sampling error increases.

Despite these limitations, simple random sampling has been preferentially used for assignment of elements randomly to different experimental conditions in psychological experiments as well as in raising a random sample for generalizing the obtained findings.

STRATIFIED RANDOM SAMPLE

In stratified random sampling the population is, first, divided into two or more strata, which may be based upon a single criterion such as sex, yielding two strata—male and female, or upon a combination of two or more criteria such as sex and graduation, yielding four strata, namely, male undergraduates, male graduates, female undergraduates and female graduates. These divided populations are called *subpopulations,* which are nonoverlapping and together constitute the whole population. Having divided the population into two or more strata, which are considered to be homogeneous internally, a simple random sample for the desired number is taken from each population stratum. Thus in stratified random sampling, the stratification of population is the first requirement. There can be many reasons for stratification in a population. Two of them are mentioned below.

1. Stratification tends to increase the precision in estimating the attributes of the whole population. If the whole population is divided into several internally homogeneous units, the chances of variations in the measurements from one unit to another are almost nil. In such a situation a precise estimate can be made for each unit and by combining all these estimates, we can make a still more precise estimate regarding the population.

2. Stratification gives some convenience in sampling. When the population is divided into several units, a person or group of persons may be deputed to supervise the sampling survey in each unit. Or, the possibility is that the institution conducting the sampling survey may have field branches to supervise the survey in each part or unit of the population.

Stratified random sampling is of two types.

(A) Proportionate stratified random sampling

(B) Disproportionate stratified random sampling

A discussion of these two types of sampling plans is given below.

(A) Proportionate Stratified Random Sampling

As its name implies, in this sampling plan the researcher stratifies the population according to the known characteristics of the population and, subsequently, randomly draws the individuals in a similar proportion from each stratum of the population. Suppose, the investigator knows the classwise distribution of population of the students in a university as given below (Table 14.2).

Table 14.2 Classwise distribution of a population of 10,000 students

Class	Composition of population	Proportion of each class
IA	$N_1 = 3{,}000$	0.30
BA	$N_2 = 4{,}000$	0.40
MA	$N_3 = 2{,}000$	0.20
M Phil.	$N_4 = 1{,}000$	0.10
	$N_t = 10{,}000$	1.00

Now, if the investigator has decided to draw a sample of, suppose, 1,000, he will include the number of students from each stratum in similar proportion. Thus, for drawing individuals from each stratum of the population in a proportionate way, he will go through a procedure as under:

Number of students from IA Class $= 1000 \times .30 = 300$
Number of students from BA Class $= 1000 \times .40 = 400$
Number of students from MA Class $= 1000 \times 20 = 200$
Number of students from M Phil. Class $= 1000 \times .10 = 100$

$$\Sigma = 1000$$

The proportionate breakdown of a sample of 1,000 students has been shown in Table 14.3.

Table 14.3 A proportionate breakdown of a sample of 1,000 students

Class	Composition of a population	Proportion of each class
IA	$N_1 = 300$	.30
BA	$N_2 = 400$	.40
MA	$N_3 = 200$	.20
M Phil.	$N_4 = 100$	.10
	$N_t = 1{,}000$	1.00

Having decided the proportion of each stratum in the sample in proportion to the population stratum and the number of students to be included in each stratum, the investigator goes on randomly drawing the students in each stratum until the desired sample size is reached. This is easily done with the help of either a table of random numbers or any suitable technique of randomness.

Proportionate stratified random sampling has many research applications. It is readily and easily applied in military studies where ranks are likely to affect other variables in the research project. Likewise, such sampling is considered important in the study of school systems and administrative services where the population remains divided into many strata. To make the sample representative of the population, it is essential to have a proportionate representation.

Proportionate stratified random sampling has some advantages and disadvantages. The important advantages are given below.

1. Proportionate stratified random sampling increases the representativeness of the sample drawn. The representativeness is strengthened because of the fact that this method of sampling ensures that those elements that exist in a few numbers are also included proportionately in the sample.

2. In proportionate stratified random sampling, the sampling error is minimized because the sample drawn possesses all the necessary characteristics of the parent population.

3. Proportionate stratified random sampling eliminates the necessity of weighing the elements according to their original distribution in the population. Since the sample is proportionately stratified according to several salient features, the sample is weighted automatically. The weight of any particular value is the frequency of it in the population.

Despite these advantages, the proportionate stratified random sampling has some limitations or disadvantages as indicated below.

1. Proportionate stratified random sampling is a difficult method. This method assumes that the researcher knows the composition and distribution of population well before the actual sampling starts. This is often an unrealistic assumption.

2. This method is a time-consuming method because the samples are drawn according to the proportion of each stratum in population.

3. This method has the probability of classification error. Since the investigator is required to identify several strata, he is likely to classify an element or individual of the population in such a way as to put him into the wrong strata. This will affect the validity and reliability of the obtained results.

(B) Disproportionate Stratified Random Sampling

Disproportionate stratified random sampling bears similarity with the proportionate stratified random sampling. The only difference is that the substrata of the drawn sample are not necessarily distributed according to their proportionate weight in the population from which they were randomly selected. In fact, some of the strata of the population may be overrepresented or some underrepresented. Sampling disproportionately means that

(a) either the investigator will give equal weight to each of the substrata, or

(b) he will give greater representation to some substrata and not enough weight to other substrata of the population in the sample to be drawn.

Suppose the investigator divides a given population of 10,000 individuals into 6,000 males and 4,000 females. If he has decided to draw a sample of 1,000 individuals from the set of 10,000 and if he draws randomly both the males and the females in equal number, say, 500 each, it will constitute the example of a disproportionate stratified random sampling. But if he randomly draws 600 males and 400 females in his sample, it will constitute the example of proportionate stratified random sampling. From this example it becomes obvious that in the disproportionate stratified random sample the investigator tries to give equal weight to each stratum, that is, he tries to draw equal number of individuals from each stratum. In doing so, he overrepresents one stratum while underrepresents the other strata. In this example, when he randomly draws 500 males and 500 females, he is overrepresenting a female stratum and underrepresenting a male stratum.

Disproportionate stratified random sampling has some advantages and disadvantages. Its major advantages are indicated as follows:

1. Disproportionate stratified random sampling is comparatively less time-consuming than proportionate stratified random sampling because here the investigator is not worried about making proportionate representation of each stratum of the population.

2. In disproportionate stratified random sampling, the investigator is able to give weight to the particular groups of elements that are not represented as frequently in the population as compared with other elements.

The major disadvantages of disproportionate stratified random sampling are given below:

1. In this method of sampling, certain stratum of the population is overrepresented and some other strata are underrepresented in the samples drawn. The assignment of greater weight to one set of elements or overrepresenting one stratum of the population may introduce some bias in the sample. Such type of sample may not be a true representative.

2. Disproportionate stratified random sampling assumes that the investigator knows the composition of the original population. This automatically means that this method of sampling cannot be used in a situation where the investigator has no idea about the composition of the original population.

3. In this method of sampling the investigator is required to classify the population into different substrata and from each substratum, he takes more or less, equal number of cases. In drawing a sample from such a classified set, he may actually misclassify certain elements, that is, he may put element A into stratum X whereas A might belong to stratum Z.

Despite these limitations, the use of disproportionate stratified random sampling is common in social science researches. In fact, because of ease and convenience it is more popular than proportionate stratified random sampling.

AREA (OR CLUSTER) SAMPLING

Area (or cluster) sampling is another important method of probability sampling. Such sampling method has its origin in the field of agriculture. Farming experiments that were conducted to determine the effect of various kinds of fertilizers, soil treatments and a variety of planting methods on crop yield, mostly used this method of sampling. In social sciences, application of area sampling has been extensive in survey research and field research.

In area sampling, generally, geographical divisions of territory, community, neighbourhood, cities, states, etc., are made on a map and a certain number of them is drawn at random and called sample. The investigator or interviewer proceeds to interview all elements of the randomly drawn areas or clusters. That is the reason why this method of sampling is also known as *Cluster sampling*.

Suppose the investigator wants to assess the attitude of the people of Tamil Nadu towards family planning. For this, it will be convenient for the investigator to have the map of Tamil Nadu before him and then divide it into various sections according to a number of vertical and horizontal grid lines drawn across the total area. He will then number each section from 1 to N, N being equal to the total number of sections. With the help of the table of random numbers, he will draw a specified number of sections to constitute the sample that he will finally study. The investigator will, then, interview all persons or members of families living in those sections. If any drawn section contains extremely different types of families, again, a random selection from among those families can be done and finally interviewed. In this way, according to necessity, further subdivision and selection of samples can be done at different stages. This is called *multi-stage sampling*.

As given hereunder, area sampling has many research utilities or applications.

(i) Large-scale surveys of political, religious and social behaviour are easily conducted by area sampling.

(ii) Where somehow lists of specific individuals are unobtainable or are inaccessible, area sampling becomes the best method of sampling.

(iii) Public opinion polls are easily and smoothly conducted using area sampling.

Area sampling has some advantages and disadvantages. The important advantages are given below.

1. When larger geographical areas are to be covered, it is easier to use area sampling than any other method of probability sampling. It is easier in the sense that the investigator need not have the list of the individuals inhabiting a given area. He simply draws some geographical sections randomly; and, subsequently, he interviews all the families or persons living in the randomly drawn sections. In this sense, area sampling is easier than the other two preceding methods of probability sampling.

2. In area sampling, respondents can readily be substituted for other respondents within the same random section. This is permitted because clusters of elements are sampled and not the individuals.

3. Area sampling saves both time and money. The investigator can concentrate his efforts in one specific region and, thus, can save time. The cost of such sampling is much less as compared to the other methods of probability sampling. The investigator need not travel great

distances to interview specific individuals residing at random points in a certain geographical region.

4. Area (or cluster) sampling possesses the trait of flexibility. In a multi-stage area sampling design, the investigator can successfully employ different forms of sampling in several successive stages. For example, after drawing several sections randomly in the first stage, the investigator may decide to stratify each of the first-stage units and select sections from some strata to constitute the second-stage units. The investigator may use a third type of sampling plan in the third or fourth stage.

5. Still another advantage of area sampling is that the respondents can readily be substituted for other respondents within the same random section. This further increases the degree of flexibility in the area sampling.

Area sampling has some disadvantages too. The important disadvantages are given below.

1. In area sampling, the degree of sampling error is usually high. This makes the sampling much less fruitful or dependable in comparison to other methods of probability sampling.

2. In area sampling, there is no correct way to ensure that each sampling unit included in an area sample will be of equal size. In fact, here the researcher has little control over the size of each cluster. This introduces bias into the samples.

3. In area sampling, it is also difficult to ensure that the individuals included in one cluster are independent of other randomly drawn clusters. It may be that an individual is interviewed in one cluster and next morning he travels to another area that falls within another randomly selected section.

Despite these disadvantages, area sampling is a common and popular method of sampling in behavioural researches.

QUOTA SAMPLING

Quota sampling is one of the important types of nonprobability sampling methods which is apparently similar to stratified random sampling. In quota sampling the investigator recognizes the different strata of population and from each stratum, he selects the number of individuals arbitrarily. This constitutes the quota sample.

Suppose the investigator knows that the population of the individuals he is going to study has three strata in terms of SES (socio-economic status)—high, middle and low. Further, suppose he knows that there are 1,000 people in high SES, 7,000 people in middle SES and 2,000 people in lower SES. Thus, the population consists of 10,000 individuals. If he wants to select 1,000 individuals and, finally, selects 100 individuals from high SES, 700 from middle SES and 200 from lower SES, according to his convenience (and not randomly), this constitutes quota sample.

It is obvious that there is similarity between quota sampling and proportionate stratified random sampling with the only difference that in the former the final selection of individuals is not random, whereas in the latter the final selection of individuals is random. Keeping in view this similarity, quota sampling is sometimes referred to as '*the poor man's proportionate stratified sample*', because such sampling seeks to ensure that specific elements are included and represented in subsequent collection of them.

Quota sampling has some advantages and disadvantages. The major *advantages* of quota sample are stated below.

1. Quota samples are the most satisfactory means when quick and crude results are desired.

2. This method of sampling is convenient and less costly than many other method of sampling, whether probability or nonprobability.

3. Quota sampling, to a greater extent, can guarantee the inclusion of individuals from different strata of population.

However, quota sampling has some *disadvantages* also as given below.

1. In quota sampling, there is no means of establishing randomness. As such, the selected samples remain no longer representative of the population. The conclusion, therefore, lacks external validity or generalizability.

2. In quota sampling, the investigators or interviewers get ample opportunity to select the most accessible individuals influencing their friends and relatives. Such readily accessible individuals may not be typical of the population they are going to study.

3. Quota sampling is amenable to classification error. Here the interviewer or the investigator bases his classification of respondents on the way they apparently look to him. In fact, he possesses no knowledge concerning the way respondents should be classified. The investigator remains ignorant of many important variables that might otherwise be used in classifying them. All these tend to make quota sampling less dependable and reliable.

4. In quota sampling, the researcher, to a greater extent, controls one variable such as sex or caste, etc., but he can't control other variables that may have both theoretical or practical significance. This mars the dependability of quota sampling.

Despite these limitations, quota sampling is a popular method among nonprobability methods of sampling, because it enables the researcher to introduce a few controls into his research plan.

PURPOSIVE OR JUDGEMENTAL SAMPLING

Purposive sample, a kind of nonprobability sample, is one which is based on the typicality of the cases to be included in the sample. The investigator has some belief that the sample being handpicked is typical of the population or is a very good representative of the population. A purposive sample is also known as a *judgemental sample* because the investigator on the basis of his impression makes a judgement regarding the concerned cases, which are thought to be typical of the population. For studying attitudes towards any national issue, a sample of journalists, teachers and legislators may be taken as an example of purposive samples because they can more reasonably be expected to represent the correct attitude than other classes of persons residing in the country. Before the start of general elections, purposive samples are often taken in an attempt to forecast the national elections. The investigator selects the persons from those states whose election results on previous polls have approximated the actual results and thus, have been typical of the whole population.

Purposive sampling has some advantages and disadvantages. The important *advantages* are given below.

1. Since purposive sampling does not involve any random selection process, it is somewhat less costly and more readily accessible to the investigator.

2. Purposive sampling is a very convenient method of sampling as compared to other methods of nonprobability sampling.

3. Purposive sampling guarantees that those individuals will be included in the sample who are relevant to the research design. The investigator does not get such guarantee in any other methods of nonprobability sampling.

Purposive sampling has some *disadvantages* also as given below.

1. In purposive sampling, there is no way to ensure that the sample is truly random or representative of the population despite the belief in typicality of the sample by the investigator. This inhibits his ability to generalize the findings.

2. In purposive sampling, too much emphasis is placed on the ability of the investigator to assess which elements or individuals are typical of population and which are not. This leaves ample scope for introducing subjectivity in the sampling. Once he has selected the sample he assumes that errors arising from his selection method will be minimal, but actually there is no legitimate stand for arguing his position.

3. In the case of purposive sampling the inferential statistics can't be used legitimately, because, under all inferential statistical techniques, there is an assumption of randomness. This criticism also applies to other forms of nonprobability sampling method.

ACCIDENTAL SAMPLING

Accidental sampling, also known as *incidental sampling*, is another popular method of nonprobability sampling plan. It refers to a sampling procedure in which the investigator selects the persons according to his convenience. Here the sample selected consists of people who are willingly available and suitable to take part and are the members of the target population. Pennington (2002) has called such sample as *opportunity sampling*. Here the investigator does not care about including the people with some specific or designated trait, rather he is mainly guided by convenience and economy. Heiman (1995) has named it as *convenience sampling*. This is a crude method of sampling, and the investigator knows that little can be generalized from the sample thus drawn.

An investigator may take students of class X into the research plan because the class teacher of that class happens to be his friend. This illustrates accidental sampling.

Accidental sampling has some advantages and disadvantages. The important *advantages* are given below.

1. Accidental sampling is the most convenient method of sampling.

2. This method of sampling possesses the trait of economy. This method saves time, money and labour of the investigator.

However, accidental sampling has some *disadvantages* as well.

1. From accidental samples, nothing can be generalized with confidence because the samples remain no longer representative of the population.

2. In accidental sampling, the investigator gets ample opportunity to show his bias and prejudice in selecting the individuals. As such, this method of sampling is not much dependable.

3. In accidental sampling, the probability of sampling error is high. Therefore, the validity and reliability of this method are badly affected.

Despite these disadvantages, it will not be an exaggeration to say that in many psychological and sociological researches this method of sampling is frequently used.

SNOWBALL SAMPLING

Snowball sampling, which is a nonprobability sampling method, is basically indirectly sociometric. It is a process of selecting a sample using networks of friends and knowns. It is defined as having all the persons in a group or organisation identifying their friends who in turn identify their friends and associates until the researcher observes that a constellation of friendships converges into some type of a definite social pattern. Some selected behaviour is usually used as the basis of contact and/or association. Obviously, then, snowball sampling is used for obtaining an impression of informal social relations among individuals. In fact, the reality is that sometimes the researcher seeks to study a 'hidden' population which is not easily identifiable, as when he deals with drug addicts, hardened criminals, prostitutes, and so on. In such cases, snowball sampling is used. Here the researcher identifies one potential subject and from him, he obtains the name and addresses of other subjects and from them, he obtains still other potential subjects so that the growth of the sample tends to build or 'snowball'. Coleman et al. (1957) used the snowball sampling successfully in a study of diffusion of medical information among physicians. The purpose was to determine how physicians come to use certain medical products such as drugs and related supplies. Snowball sampling revealed the pattern of diffusion of information. Besides, such sampling can also locate the physicians' cliques through which such information is revealed. It may answer questions such as: Do the physicians read about new drugs in journals or hear about it in some medical seminar? If they hear about it in some medical seminar or convention, whom do they contact among their physician friends regarding it? How does information about a given drug spread among physician fellows?, and so on.

Snowball sampling has important research applications in relatively small business and industrial organisations where N is expected not to exceed 100. Such sampling is more convenient to the studies of social change and diffusion of information among specific segments of social organizations.

Snowball sampling has some advantages and disadvantages. The important *advantages* are given below.

1. Snowball sampling, which is primarily a sociometric sampling technique, has proved very important and is helpful in studying a small, informal social group and its impact upon formal organizational structure.

2. Snowball sampling reveals communication pattern in community organization concepts like community power; and decision-making can also be studied with the help of such sampling technique.

3. The method of snowball sampling is amenable to various scientific sampling procedures at various stages such as use of random numbers or computer determination.

Despite these advantages, snowball sampling has some *limitations* also as described below.

1. Snowball sampling becomes cumbersome and difficult when N is large or say it exceeds 100.

2. This method of sampling does not allow the researcher to use probability statistical methods. In fact, the elements included in sample are not randomly drawn and they are dependent on the subjective choices of the originally selected respondents. This introduces some bias in the sampling.

Despite these limitations, snowball sampling is used in the study of diffusion of some specific information and social changes.

SATURATION SAMPLING AND DENSE SAMPLING

Coleman (1959) has emphasized these two types of sampling techniques which are used less frequently as compared to other techniques of sampling. Saturation sampling is defined as drawing all elements or individuals having characteristics of interest to the investigator. Drawing all physicians having at least the age of 45 (from a small community) would be called saturation sampling. Dense sampling is a method of sampling which lies somewhere between simple random sampling and saturation sampling. When the researcher selects 50% or more from the population and takes a majority of individuals having specified traits or characteristics which are of interest to him, it is called dense sampling. For example, if the researcher selects 500 to 600 students from a population of 1,000 students all having distinction marks in any one examination paper, it will constitute dense sampling.

The advantage of saturation sampling and dense sampling is that these are convenient methods of sampling where N of the population does not exceed 1,000. In case it exceeds, these methods of sampling become cumbersome and inconvenient.

DOUBLE SAMPLING

Double sampling, as its name implies, is defined as drawing a sample of individuals from another sample of them. Suppose the investigator randomly draws a sample of 1,000 from a population having $N = 10{,}000$. From this 1,000 individuals, he again randomly draws a sample of 300 for further study. This is called double sampling.

Suppose the researcher is studying the attitude of newly married couples towards family planning through mail questionnaire. For this, he mailed 1,000 questionnaires to one thousand couples residing in different localities. He finds that only 40% (that is 400) questionnaires were returned. From these 400 persons, he draws randomly a sample of 100 and mails another questionnaire to get their in-depth knowledge towards the different techniques of family planning. This is called double sampling.

For double sampling to be an effective part of the research, it is essential that the sampling plan followed from the beginning should maximize the representativeness of the individuals drawn subsequently. This method has the disadvantage of taking much time and labour of the researchers.

MIXED SAMPLING

Mixed sampling is one where the characteristics of both probability sampling and nonprobability sampling are mixed. Systematic sampling is one such sampling.

Systematic sampling may be defined as drawing or selecting every *n*th person from a predetermined list of elements or individuals. Selecting every 5th roll number in a class of 60 students will constitute systematic sampling. Likewise, drawing every 8th name from a telephone directory is an example of systematic sampling. If we pay attention to the systematic sampling plan, it becomes obvious that such a plan possesses certain characteristics of randomness (first element selected is a random one) and at the same time, possesses some nonprobability traits such as excluding all persons between every *n*th element chosen.

Systematic sampling has some advantages and disadvantages. The important *advantages* are mentioned below.

1. Systematic sampling is relatively a quick method of obtaining a sample of elements. If the investigator has a short time schedule, this method of sampling eliminates several steps otherwise taken in different methods of sampling.

2. Systematic sampling makes it very easy to check whether every *n*th number or name has been selected. In case there occurs any error in counting, that is, if the investigator selects 6th number instead of 5th number, his sample will not be seriously affected.

3. Systematic sampling is very easy to use. In fact, it is much simpler than having to employ a table of random numbers for drawing the sample or fixed quota from each stratum of the population in order to have proportional representation.

Despite these advantages, there are some *limitations* of systematic sampling as indicated below.

1. Systematic sampling ignores all persons between every *n*th element chosen. Obviously, then, it is not a probability sampling plan.

2. In systematic sampling, the sampling error increases if the list is arranged in a particular order, say, the list increases or decreases with respect to some trait such as age, education, income, caste, etc. In such a situation, a bias will be introduced in the sample because more individuals are likely to be drawn from one group, particularly from that group which is represented most frequently on the group. Such type of bias may also result when the list is alphabetical (Blalock 1972).

Despite these limitations, systematic sampling is frequently used in psychological and sociological researches because this method of sampling possesses the trait of both probability and nonprobability sampling.

REQUISITES OF A GOOD SAMPLING METHOD

Any sampling method to be good and scientifically sound must possess at least the two following properties:

1. It must ensure the representativeness of the sample.
2. It must ensure the adequacy of the sample.

The first requirement of a sampling method is that the selected samples must be representative of the population. What is meant by representativeness of the sample? In the words of Kerlinger (1973, 119), "Ordinarily, representative means to be typical of a population, that is, to exemplify the characteristics of the population... . In research a representative sample means that the sample has approximately the characteristics of the population relevant to the research in question." Truly speaking, an investigator cannot be certain about the representativeness of a sample unless the entire population is tested. But this is ordinarily not feasible. Hence, he must satisfy himself otherwise. A sample can be regarded as being representative of the population when it possesses all the relevant characteristics of the population in about the same proportion as is in the population. If sex is the relevant characteristic of a population of students of a particular university and if the population has a male:female ratio of 60:40, a representative sample would be one where the ratio of males and females is also 60:40. For ensuring representativeness of samples it is also essential that the population should be clearly defined and a suitable and appropriate definition of the observations, which are to constitute samples, be framed. An observation must be defined in the light of the definition of population; otherwise the result will be a biased sample. Suppose that the population is defined as all those females who are widows. Further, suppose that the observation is defined in terms of selection of those who admit that they are widows. In this case, the resulting samples will not be representative ones because observation has been defined differently from the population. Obviously, this definition of observation will automatically exclude those who refuse to admit that they are widows and therefore, the sample will be a biased one.

The second requirement of good sampling is that the samples must be adequate. A sample is said to be adequate when it is of sufficient size. There is no rigid rule regarding the size of a sample. But a larger sample, in general, is preferred because it tends to reduce the error, which is the difference between the population value and the sample value. The mean of a sample taken from a population will not ordinarily be equal to the mean of that population. The relationship between the size of the sample and the error, as defined above, is negative. The larger the size of a sample, the smaller the error and vice versa. The negative relationship between the size of a sample, and the error is demonstrated in Figure 14.2. When the size of the sample is increased, the form and the statistics of the sample distribution become a close approximation to those of population. Hence, the error is reduced and therefore, a large sample is preferred. But a very large sample may not yield that result because it tends to create several problems, which sometimes become difficult to solve.

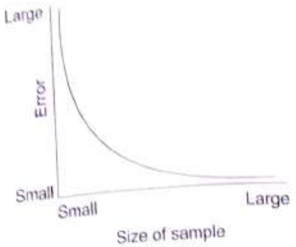

Fig. 14.2 Relationship between the size of sample and the error

COMMON ADVANTAGES OF SAMPLING METHODS

The purpose of any sound sampling procedure is to make sampling more efficient and easy. In any behavioural research, the whole population is not examined and an estimate of the characteristics of population is made on the basis of the characteristics of the sample drawn from the population. Investigators rely upon samples and the conclusion is generalized. Apart from serving this function of generalization, sampling has some additional advantages.

1. *Accuracy is increased*: When one is entrusted with examining sample (and not the population), naturally, a great volume of the work is reduced. Consequently, one can easily supervise the task assigned to the sample and the results can also be analyzed with great care. Not only this, the services of trained personnel can be utilized at a lower cost when the number of cases to be processed is low. The overall results of all this is that the accuracy of the investigation in increased.

2. *Cost is reduced*: When the data are to be collected from a limited number of cases rather than from a larger population, this naturally entails a reduced cost to the investigator. Not only this, the time taken in processing of the data is also reduced. When one is dealing with only a limited number of data, the chances of accidental error in statistical calculation are also minimised.

3. *Speed is increased*: When the data are collected from a sample rather than from the entire universe, the investigator can do his work speedily without doing any injustice to the investigation. This is one of the vital considerations, specially when the information about the sample is needed urgently.

SAMPLING DISTRIBUTION

In any sampling situation, three types of distributions are of common interest to the research investigator: *the sample distribution*, *the population distribution*, and *the sampling distribution*. The meaning of and distinction between these three types of distributions may be explained through an illustration.

Suppose in a small town, the total number of graduate parents is 30,000. Suppose further that we are interested in knowing the mean fertility rate among these parents. Owing to time and money considerations, we cannot examine the whole population and therefore, we decide to examine only a random sample of 300 parents. The observed data may be arranged in a frequency distribution and the desired statistics may be calculated. This type of distribution is known as the *sample distribution* because it shows the distribution of data obtained from a sample drawn randomly from the population. Now, if with the help of any social organization or institution we are able to record the fertility rate of 30,000 parents and arrange them in frequency distribution, this distribution is termed as the *population distribution*. In reality, the forms and the parameters of a population are not ordinarily known. Therefore, an estimate about these two properties of the population is made on the basis of the form and parameters of the sample distribution. If the form of sample distribution is normal, we expect that the form of the population distribution will also be normal. If the mean and the standard deviation of the fertility rate of the 300 graduate parents are 3.89 and 1.02 respectively, we assume that the mean and the standard deviation of the population distribution are also 3.89 and 1.02 respectively. If we take all possible samples of a given size (in this example, 100 samples consisting of 300 parents each can be taken), we will get different sample distributions. For illustration, we will get 100 different sample distributions, that is, 100 different means and 100 different standard deviations. If we group the 100 means and the 100 standard deviations into frequency distributions separately, we will get the sampling distribution of mean and the sampling distribution of standard deviation. Thus, the sampling distribution is a *theoretical* or *idealized distribution*, which is obtained if all possible samples of the given size are taken from the population. Thus a sampling distribution is a distribution of statistics obtained by selecting all the possible samples of a specific size from a population.

SAMPLING ERROR

As the reader must have gathered by now, sampling refers to the process of drawing a limited number of cases from a population. Whatever the sampling procedure used, the sample must represent, as far as possible, all characteristics of the population that are related to the problem under study.

A statistical measure calculated from a sample is known as a statistic and the corresponding measure, which is usually unknown and is based directly upon the population, is known as a parameter. A statistic is frequently referred to as an obtained measure or 'derived measure'. A mean calculated from a sample of a population is called 'statistic mean' and the value of mean based directly upon the population is known as 'parameter mean'. A sampling error is the difference between the parameter and the statistic. As Lindquist (1968, 8) has put it, "A sampling error is the difference between a parameter and an estimate of that parameter which is derived from a sample."

A sampling problem is said to exist whenever the conclusions obtained on the basis of a limited number of individuals (called the sample) are applied to a larger number of individuals (called the population). Such generalized conclusions are known as *statistical inferences*. The higher the sampling error, the poorer the statistical inference. The term 'sampling error' does not suggest any mistake in the process of sampling itself, rather it merely suggests the chance variations which are inevitable when a number of randomly selected samples are taken from the same population. Due to this sampling error, making an estimate about the population characteristics from the sample characteristics no longer remains an exact process. It has been observed that a sample statistic does not ordinarily agree with that obtained from the second, third or fourth sample or with the population parameter. The amount of fluctuation (or the sample error) depends upon the following two factors:

(i) Variability in the population

(ii) Size of the sample

If the nature of the population is such that it is not stable and tends to vary, naturally, the successive samples taken from the same population over time will tend to differ. Generally, a random sample provides a good approximation to the population and the accuracy of approximation depends upon the size of the sample. As the size of the sample is increased, the form and the statistics of the sample distribution come nearer to those of the population. For example, when the size of the sample is 100, the distribution as represented by the frequency polygon is usually smooth enough to give a very clear idea about the form of the population distribution.

There are certain statistics which tend to fluctuate less than others in successive samples and hence, the sampling error is less. For example, in sampling from a normal distribution the mean and the standard deviation fluctuate less than other measures of central tendency and measures of variability. Hence, these two are more reliable than other similar measures because they approximate more closely the corresponding population parameters and hence, tend to fluctuate less from sample to sample and thus produce the least sampling error.

In most sampling situations, there are three types of important distributions which are of concern to the reader in order to have a clear-cut idea about the application of the sampling error. They are: the sample distribution, the population distribution and the sampling distribution. These three terms may be explained with an example. Suppose, we wish to know the mean intelligence score of workers (in the age range of 20–45 years) in the Tata Iron and Steel factory. Further suppose that the total number of workers in this given age range is 30,000. Owing to time and money considerations, we may decide to take a random sample of 300 and administer a standardized test of intelligence. The obtained scores on the test by 300 workers expressed in frequency distribution will be called the sample distribution. Thus, the sample distribution is an observed distribution which can be described by any descriptive statistics. Hence, sample distribution is the concrete distribution in any sampling situation.

However, if we decide to administer the test of intelligence on the whole population, that is, on 30,000 workers, we shall be getting a set of 30,000 scores. If we convert this set into frequency distribution, this will be known as the population distribution. Ordinarily, the form and the parameters of the population distribution are not known. However, we can infer them on the basis of the sample distribution. If the sample distribution is normal, it can be inferred that the population distribution is also normal, ascribing any irregularities in the sample distribution to the sampling fluctuations.

The inferences drawn regarding population distribution on the basis of the sample distribution may not be correct due to sampling errors. In the above example, we do not have the fixed mean and the fixed standard deviation which would be obtained again in the second or third sample of 300 from the population. If we were to take, say, 100 samples of 300 each, we would be getting 100 different means, 100 varying standard deviations. If all the 100 means and 100 standard deviations are grouped in frequency distribution separately, we would have an exact sampling distribution of mean and the standard deviation in the population. As this is not done, sampling distribution is a theoretical or idealized distribution which may be thought of as a distribution which would be obtained if the values of a given statistic were computed in all possible numbers of a sample of given size actually taken from the population.

Statisticians have been able to determine the sampling distribution of the majority of commonly used statistics in the samples taken from the normal population. When the sampling distribution of a statistic is known, it is possible for the investigator to determine the relative frequency with which different sample values tend to occur in random sampling from a population in which the statistics have some of them assumed or hypothesized. After knowing the relative frequency, the investigator is able to know the probability that a particular sample value has resulted from the sampling fluctuations or chance. Ultimately, by knowing the probability figure, he will be able to judge the soundness of a hypothesis regarding the value of the statistics in a population.

☑ Review Questions

1. What do you mean by sampling? Discuss the factors that influence the decision to sample.
2. What is sampling distribution? Discuss the methods of drawing random samples.
3. Citing some examples, discuss the relative advantages and disadvantages of any two probability sampling methods.
4. Citing some examples, discuss the relative advantages and disadvantages of any two nonprobability sampling methods.
5. What is 'systematic sampling'? How does it differ from 'cluster sampling'? Discuss the relative advantages and disadvantages of systematic sampling.
6. Explain the importance of sampling in research. Discuss any two methods of sampling.
7. What are the requisites of a good sample? State the advantages of randomization in scientific research.
8. What is the greatest danger in using accidental sample in research?
9. When is cluster sampling likely to be preferred most?
10. Why would a researcher choose stratified random sampling instead of simple random sampling?

◆

15

SOCIAL SCIENTIFIC RESEARCH

CHAPTER PREVIEW

- Meaning and Characteristics of Scientific Research
- Scientific Approach to the Study of Behaviour
  - The Assumption of Science
  - Attitudes of Scientists
  - The Flaws in Scientific Research
- Validity in Research or Experimental Validity
- Controlling Threats to Reliability and Validity in Research
- Phases or Stages in Research
  - Identifying the Problem
  - Formulating a Hypothesis
  - Identifying, Manipulating and Controlling Variables
  - Formulating a Research Design
  - Constructing Devices for Observation and Measurement
  - Summarizing Results
  - Carrying out Statistical Analysis
  - Drawing Conclusions
- Types of Educational Research
- Types of Research: Experimental and Nonexperimental
  - Laboratory Experiments
  - Field Experiments
  - Field Studies
  - Difference between Field Experiment and Field Study
  - Ex Post Facto Research
  - Survey Research
  - Document or Content Analysis
  - The Case Study
  - Ethnographic Studies
- Difference between Research Method and Research Methodology
- Ethical Problems in Research
- Comparison between Experimental Research and Nonexperimental Research
- Types of Experiment
- Types of Applied Research

MEANING AND CHARACTERISTICS OF SCIENTIFIC RESEARCH

The word research consists of two syllables— 're' and 'search'. The dictionary meaning of the prefix 're' is again or over again and the dictionary meaning of the word search as a verb is to examine carefully or to probe. Together they form the noun research describing a careful,

systematic investigation in some field of knowledge for establishing facts or principles (Grinnell 1993). Therefore, any scientific research is a systematic and objective attempt to provide answers to certain questions or problems. The purpose of scientific research is to discover and develop an organized body of knowledge. Therefore, scientific research may be defined as the systematic and empirical analysis and recording of controlled observation, which may lead to the development of theories, concepts, generalizations and principles, resulting in prediction and control of those activities that may have some cause-effect relationship. A similar definition has been given by Kerlinger (1973, 11): "Scientific research is a systematic, controlled, empirical and critical investigation of hypothetical propositions about the presumed relations among natural phenomena." In order to elaborate the above views, some of the characteristics of systematic research are presented below.

1. Research is always directed towards the solution of a problem. In other words, in research the researcher always tries to answer a question or to relate two or more variables under study.

2. Research is always based upon empirical or observable evidence. The researcher rejects those principles or revelations, which are subjective and accepts only those revelations, or principles, which can be objectively observed.

3. Research involves precise observation and accurate description. The researcher selects reliable and valid instruments to be used in the collection of data and uses some statistical measures for accurate description of the results obtained.

4. Research gives emphasis to the development of theories, principles and generalizations, which are very helpful in accurate prediction regarding the variables under study. On the basis of the sample observed and studied, the researcher tries to make sound generalizations regarding the whole population. Thus, research goes beyond immediate situations, objects or groups being investigated by formulating a generalization or theory about these factors.

5. Research is characterized by systematic, objective and logical procedures. The researcher tries to eliminate his bias and makes every possible effort to ensure objectivity in the methods employed, data collected and conclusion reached. He frames an objective and scientific design for the smooth conduct of his research. He also makes a logical examination of the procedures employed in conducting his research work so that he may be able to check the validity of the conclusions drawn.

6. Research is marked by patience, courage and unhurried activities. Whenever the researcher is confronted with difficult questions, he must not answer them hurriedly. He must have patience and courage to think over the problem and find out the correct solution.

7. Research requires that the researcher has full expertise of the problem being studied. He must know all the relevant facts regarding the problem and must review the important literature associated with the problem. He must also be aware of sophisticated statistical methods of analyzing the obtained data.

8. Research is replicable. The designs, procedures and results of scientific research should be replicable so that any person other than the researcher may assess their validity. Thus, one researcher may use or transmit the results obtained by another researcher. Thus, the procedures and results of the research are replicable as well as transmittable.

9. Finally, research requires skill of writing and reproducing the report. The researcher must know how to write the report of his research. He must write the problem in unambiguous terms; he must define complex terminology, if any; he must formulate a clear-cut design and procedures for conducting research; he must present the tabulation of the result in an objective manner and also present the summary and conclusion with scholarly caution.

SCIENTIFIC APPROACH TO THE STUDY OF BEHAVIOUR

As we know, the major difference between scientists and nonscientists lies in the way scientists go about answering the questions they ask about nature. Scientists use *scientific methods*. The scientific method consists of some assumptions, attitudes, goals and procedures for creating and answering questions about nature.

Psychology is a science like physics, chemistry or biology because they all employ the scientific method. Scientists employ a certain philosophy towards the entire issue of learning about nature. In fact, this philosophy starts with the way scientists try to conceptualize nature.

The Assumption of Science

Any aspect of nature, especially human behaviour, seems not only complex but also mysterious. Despite this, we make the effort to understand it. In so doing, science makes certain assumptions about nature that helps us to study it as a regulated and consistent system. These assumptions may be explained as under:

1. *Nature is lawful*: The basic assumption underlying the scientific method is that nature tends to operate in a lawful manner. This assumption is made because if nature is not lawful, rather is random, it can never be understood fully. As physicists discuss how the 'law of gravity' governs the behaviour of planets, psychologists also assume that there are laws of nature that govern the behaviour of living organisms. Behaviours are regulated by natural laws and therefore, any behaviour can be explained in terms of specific causes and effects. For example, when psychologists are studying the development of language in humans, they are, in fact, studying the laws of nature.

Some psychologists, however, would argue otherwise, claiming that in fact, it is too much to say that a research study directly examines a law of nature because of a narrow and limited focus of each study. Apart from this, most researchers are frequently confronted with contradictory findings and opposing explanations. However, we can assume that eventually all of the diverse findings will be integrated and coordinated so that psychologists can truly understand the laws of nature.

2. *Behaviour is deterministic*: Psychologists assume that the behaviour of the organism is determined. According to the doctrine of determinism, behaviour is solely influenced by natural causes. Determinism, which is distinct from predestination, suggests that while there is no overall plan, there are predictable, identifiable, and natural causes for every behaviour. Thus behaviour of the living organism does not result from 'free will' or choice. Thus the science of psychology clearly states that persons cannot freely choose to exhibit a particular personality or respond in a particular way in a given situation. The laws of behaviour force persons to exhibit certain attributes and to behave in a certain way in a certain situation. Any two persons in the same situation will be similarly influenced because that is how the laws of behaviour operate. But due to some individual differences, no two individuals will behave identically.

3. *Nature is understandable*: Although nature is mysterious and complicated, we assume that it can be understood. We may have not yet fully understood some aspects of nature, but it is assumed that we will eventually be able to understand it. This assumption has two important implications. *First*, any statement cannot be taken on faith. In science, it is always appropriate to question and to ask for proof. *Second*, any scientific statement must logically and rationally fit within the known facts so that it can maximally be understood.

Conclusively, it can be said that any behaviour, in order to be studied scientifically, must be assumed to be *lawful*, *determined* and *understandable*.

Attitudes of Scientists

Scientists tend to differ from nonscientists in their attitudes towards understanding nature. They employ certain types of attitudes that guide their activities in a certain way.

1. *Scientists are open-minded*: Scientists adopt an attitude of open-mindedness in their approach. They must leave their biases and preconceptions behind. Any explanation may offend their sensibilities or contradict their beliefs, but simply due to such contradictions, they don't dismiss it. They look in different directions, at all possible aspects, when trying to understand a behaviour.

2. *Scientists are uncertain*: Scientists recognize this basic principle that there is always a degree of uncertainty. At any one point they may feel that they have some degree of understanding of nature, but no one already knows everything about how nature operates. For psychologists, this obviously means that no one knows precisely what a behaviour entails, no one knows the appropriate and correct way to study a behaviour, and no one knows all of the factors that influence behaviour.

3. *Scientists are fallible*: Scientists develop an attitude that everyone is fallible. We make mistakes because we don't know how nature works. If we pay attention to the history of science, it will be obvious that the description that at first appeared accurate later turned out to be wrong or inaccurate. Therefore, scientists always question whether the factors proposed as important might actually be irrelevant, and whether the factors proposed as irrelevant might actually be important.

4. *Scientists are cautious*: Scientists are very cautious when dealing with scientific findings. They never treat the results of any single study as a fact in the usual sense. Instead, they treat a research as finding merely a piece of evidence that provides them with some degree of confidence in the statement about nature. For this reason, the researcher should never use the terms 'proof' or 'proved'.

5. *Scientists are skeptical*: Scientists are very skeptical about accepting the truth of any scientific statement. This is because of the fact that no one knows how nature works, and because we are fallible, any description of nature may prove to be incorrect. Scientists critically evaluate the evidence using various types of logic and their knowledge of psychological researches in order to know whether the evidence for any statement is correct or conversely, is fraught with several kinds of problems and contradictions. There is also another reason for scientists to be skeptical: Scientists sometimes do fraudulent acts, that is, they may publish research findings when no research was actually conducted or sometimes inaccurately report their findings. Often in such a situation, they simply try to respond to professional pressures to be a productive researcher.

6. *Scientists are ethical*: Scientists adopt an ethical attitude as they are very much concerned about balancing the goals and desires of the researcher with responsibility to cause no harm to others. *First*, researchers in psychology have an ethical responsibility to conduct a research in such a way that may not cause harm to others. *Second*, researchers have also ethical responsibility of not conducting any type of fraudulant act, which includes falsifying results and keeping such a result secret that contradicts one's views.

In this way, it can be said that scientists have adopted certain attitudes towards nature including conduct of human behaviour. They recognize fallibility of all researches related to it. Therefore, they are open-minded, skeptical, cautious and ethical, too, in dealing with subjects or participants of the study.

The Flaws in Scientific Research

Any scientific research employs many rules and objective criteria. Despite this, these rules and criteria don't guarantee to eliminate bias and error completely from the research. As we know, there are many ways to design a study. Since there are many ways to design a study, and human nature is complex, there are also many opportunities for researchers to make errors in either collecting data or interpreting it. Therefore, flaws in scientific research basically relate to *flaws in scientific evidence* and *flaws in testing hypotheses* and its interpretation. We shall consider those two flaws separately.

Flaws in Scientific Evidence

Ideally in any scientific research, observations should be completely *empirical, objective systematic* and *controlled,* so that the researcher may be accurate in observation and measurement of the exact behaviour. However, the reality is that no study is perfect and ideal. In fact, any study may be more or less empirical, more or less objective, more or less systematic and more or less controlled. Basically, there are four factors that easily determine the degree to which a study meets the criteria for scientific evidence.

1. *Nature of behaviour to be studied*: There are some implicit behaviours which cannot be studied in a completely empirical, objective, systematic and controlled way. For example, thinking is such an implicit behaviour. We cannot observe thinking directly, and we will have to observe some other behaviours and then draw inferences about thinking. Similarly, the extent to which we are able to obtain precise and objective measurements is much dependent upon the behaviour being studied. For example, for objectively assessing and measuring behaviours like aggressiveness or love, there is no yardstick and there is no single behaviour from which inferences can be drawn. For assessing these behaviours, the researchers are bound to employ some less concrete and more subjective measurement procedures which may include some biases or errors, too. Likewise, a researcher cannot observe all aspects to all behaviours in a systematic and controlled way. For example, suppose he is studying attitudes of females towards child rearing. Here it is difficult for the researcher to separate the fact that each subject has female's personality, from the fact that each subject has also a female's genes and physiology. In such situation, the researcher cannot be sure whether it is a female's personality or her physiology (or both) which are influencing her attitudes.

In view of the above limitations, flaws creep into the scientific evidences and then, we cannot know what our measurements reflect about the behaviour or what factors are truly operating.

2. *Technical limitations*: Sometimes there is a lack of necessary technical capabilities and thus the researcher is restricted in observing and measuring some behaviour or factor. In such a situation, the research may yield very limited and potentially misleading information. For example, in the late 1800s, 'phrenology' was included in psychology. As we know, phrenology basically denotes that various personality traits are reflected by the size of the parts of the brain and thus, by bumps on certain parts of the skull. Considering today's technology for studying brain physiology and personality, the study of phrenology seems silly.

3. *Properties of the research design*: The properties of the research design also tend to produce some flaw in scientific evidence and thus reduce our confidence in the findings. In fact, the research design specifies the situation where the study is to be conducted (that is, whether laboratory or field), the types of subjects we will study, the way we will measure a behaviour and the way we control or manipulate factors. All these specifications of the research design, no doubt, on one hand, increases objectivities to some extent but also tends to restrict and reduce our confidence in the findings.

4. *Limited perspective of a study*: The evidence obtained from a particular study may be convincing but it may not tell the whole story. Any one study may be considered as a 'snapshot' that necessarily provides a slanted view, considering certain factors and ignoring others. In other words, due to limited perspective, any study may not represent the nature because in one study the researcher takes one approach which automatically means that he is excluding others. In fact, this situation is illustrated by an old fable which tells of several blind men trying to describe an elephant. In studying a behaviour, the researchers are like the blind men, trying to describe the entire elephant by taking only one limited perspective at a time.

These are the major flaws in scientific evidence. Even when the evidence from a study reflects a minimum of flaws, we still cannot be sure about its findings and conclusion because the interpretation of the results is also to be taken into consideration. Inappropriable interpretation of testing the hypotheses may introduce several flaws.

Flaws in testing hypotheses

Flaws in testing hypothesis are also important sources of flaws in scientific research. Often, it is found that while testing a psychological hypothesis, some data appear to confirm a hypothesis while other data appear to disconfirm it. Therefore, it is essential that the researcher must critically evaluate and weigh the various types of evidence and research that produced it. Let us illustrate the whole sequence through an example. Suppose the hypothesis being tested is: Dreams predict future. If the person accepts it, it is probably because he has occasionally dreamed of events that ultimately occurred. But let us consider the quality and quantity of evidence that supports this hypothesis. There are several points to be considered. *First,* this hypothesis is a suspect because it is neither rational nor parsimonious. *Second,* the recall and interpretation of the dream and the feeling about how they match up with the real events actually lack objectivity. *Third,* here the researcher is relying only upon confirmation. His observation is consistent with the hypothesis that dreams tend to predict the future. But this weak evidence should be weighed against the amount of disconfirmation that is available, the number of times the dreams failed to come true. In sum, it can be said that there are many evidences which disconfirms the hypothesis that dreams predict the future. Therefore, it could be said that those few confirming instances are nothing more than coincidence.

Thus it is obvious that our confidence in any scientific hypothesis is based on the quantity and quality of evidence that confirms and disconfirms it. Accepting a hypothesis merely on the basis of confirmation and ignoring all those evidences that provide disconfirmation introduces important flaws in scientific research.

Finally, we can conclude that flaws in scientific evidence and in testing hypothesis and its interpretection provide a big challenge to modern researchers.

VALIDITY IN RESEARCH OR EXPERIMENTAL VALIDITY

Every researcher attempts to achieve maximum validity in his research work. Ordinarily, he has two basic aims in his research work. *First,* the researcher wishes to see the extent to which the factors or variables manipulted in the study have a systematic effect upon the desired variable in an experimental setting and the extent to which the desired variable is not affected by uncontrolled or extraneous factors. This aim of the researcher is called *internal validity* of the research. Internal validity is when the independent variable alone affects the dependent variable. Thus internal validity indicates the extent to which an experiment is methodologically sound and free from confounding. Researches in which the effect obtained cannot be attributed with confidence to the independent variables are called *internally invalid* research. In contrast, the research has internal validity when the effect can be confidently attributed to the manipulation of the independent variable. Tuckman (1978, 4) has said, "A study has internal validity if the outcome of the study is a function of the programme or approach being tested rather than the result of other cases not systematically dealt with in the study." Besides the above aim, the researcher has another aim, which is primarily concerned with the *generalizations* of the research. He wishes to generalize the conclusion reached on the basis of his research in relation to similar situations or programmes in nonexperimental settings. The degree to which this goal is achieved reflects the *external validity* of the research. Again, to quote Tuckman (1978, 4), "A study has external validity if the result obtained would apply to the real world, to other similar programmes and approaches." Of these two types of validity, internal validity is of high value because if a research or experiment has little or no internal validity, its conclusion cannot be generalized or can be generalized only with considerable risk and hence, its external validity will automatically be affected.

There are several factors that tend to influence the internal and the external validity of research. In behavioural experiment many extraneous variables are present, which are likely to influence the dependent variable. Extraneous variables (also known as relevant variables) are those uncontrolled variables or unmanipulated variables that may produce a significant effect upon the dependent variables. Although these extraneous variables cannot be altogether eliminated from the experimental setting, they can be recognized and then their influence upon the results of the experiment can be minimized. Campbell & Stanley (1966) have pointed out a number of extraneous variables or factors that are likely to affect the internal validity and the external validity. Within the category of internal validity are a number of such factors. Some of them are mentioned below.

1. *Maturation*: Maturational changes are likely to be confused with changes produced by the manipulation of independent variables. This is particularly true in the case of a long-term experiment. When the experiment continues for a longer period of time, the subjects may become wiser due to incidental learnings occurring during the period of exprimentation. In such a case, a change is likely to occur which may provide a threat to the internal validity of the experiment.

2. *Unsound instrumentation*: Normally, the researcher uses some instruments or tests for assessing different aspects of behaviour. If those tests are less reliable and less valid, errors are likely to be introduced in the measurement, which may ultimately tend to reduce the internal validity of the experiment. If human observers are used in place of tests or instruments, errors relating to human observations such as errors due to fatigue, increased experiences, and changes in mood of the observers are likely to affect the internal validity of the experiment.

3. *Immediate external situations*: Immediate external situations are likely to affect the internal validity of the research. The subject may be worried due to failure in a university examination; he may not have slept the previous night due to fatigue from a late night party; or he may be too shocked after hearing the news of the death of his mother or father, and so on. All these external events are beyond the control of the experimenter or the researcher and tend to reduce the subjects' consistency in behaviour and thus the internal validity of the research.

4. *Testing*: Testing means the effects of pretesting upon the performance of the subjects during the post-testing. This process of testing sometimes also tends to induce a threat to the internal validity of the research. When the subjects are tested in the beginning of the experiment, two consequences are likely to be the most obvious ones on the post-testing performance. First, the subjects may be very conscious of the latent purpose of the test. This, in turn, is likely to produce some changes within the subject. Second, the pretesting may make the subjects more skilled and proficient on subsequent testing or post-testing. This, again, is likely to produce some changes in the subjects. Both these types of changes are likely to affect the internal validity of an experiment adversely.

5. *Selection bias*: Selection bias is likely to affect the internal validity results when the researcher makes a comparison between the nonequivalent experimental and control groups. Due to one reason or the other, when the control group is not equated to the experimental group, these two groups are likely to differ in motivation, interest and even in ability. All these differences are likely to affect the internal validity adversely. Probably the best way to control this type of selection bias is to make use of the random assignment (also called randomization) of subjects to the experimental group and the control group.

6. *Experimental mortality*: Experimental mortality means the loss of subjects during the period of experimentation. Experimental mortality is often found in the case of long-term experiments. Generally, it has been reported that those subjects who remain during the entire period of experimentation are more motivated, more interested and more attentive than those who either drop out in the middle or forcefully remain present. This produces a confounding effect in the sense that it creates an imbalance among the subjects, no matter how randomly they had been initially selected and ultimately poses a threat to the internal validity.

7. *Statistical regression*: The phenomenon of statistical regression is often seen in pretest-post-test situations and where the groups are selected on the basis of extreme scores. Subjects who score higher on pretest (that is, selected on the basis of high scores) tend to score lower on post-test measurement and the subjects who score lower on pretest, tend to score higher on post-test measurement. Normally, this regression takes place towards the mean of the distribution. In other words, the scores of the high scorers will tend to decrease towards mean while the scores of the low scorers will tend to increase towards the mean on post-test measurement. Thus the two groups differ on the post-test scores even without any treatment given to them. Probably, the reason is that chance factors affect the extreme scores more than average scores. When a researcher fails to recognize this, a threat to the internal validity is apparent.

8. *History*: The term 'history' as used in research refers to those events of the environment which occur at the time the experimental variable is being tested. Suppose tomorrow a group of students (all Indians) is to be administered a test measuring prejudice towards the Chinese. The same night China invades India. Such external events are likely to affect the scores on the test of prejudice. Probably, the subjects will show a strong prejudice against the Chinese. In such a situation, the measured outcome on the test may not reflect the true prejudice.

9. *Diffusion of treatment:* Diffusion of treatment also known as *contamination* is such a threat that research participants in different groups are likely to communicate with each other and learn about the treatments of others. Experimenters try to avoid it by isolating groups or let the participants promise not to reveal anything to others who will act as participants. For example, suppose an experiment is being conducted in which the participants are kept engaged for the whole day on a new way to memorize the task. During break, the participants of the experimental group tell those participants in the control group about the new way to memorize and that technique may be used by the participants of the control group. Such diffusion of treatment to the control group is likely to provide threat to the internal validity of research.

10. *Compensatory behaviour:* Sometimes experimenters may provide something of value to one group of participants but not to another. Such differences become known. This inequality tends to produce pressure to reduce differences and rivalry between groups, or sometimes resentful demoralization is also produced. These behaviours, called compensatory behaviour, can affect the dependent variable in addition to the treatment. Such effect naturally becomes confounding. For example, suppose one school system receives a longer lunch break (treatment) for producing gains in learning whereas another doesn't and rather continues on doing hard work for learning. Such system of the same school may be demoralizing and the participants or students may withdraw from learning.

11. *Experimenter's expectation:* Although experimenter's expectation is not always considered a threat to the internal validity, sometimes the experimenter's behaviour can threaten internal validity. An experimenter may threaten internal validity by indirectly communicating experimental expectancy to the participants. The experimenter may be highly committed to the problems and hypotheses of the study and indirectly communicate desired findings to the participants. For example, suppose the experimenter has framed the hypothesis that female participants will be more sensitive towards disability than male participants. Through tone of

voice, eye contact, pauses and other nonverbal communication, the experimenter may unconsciously encourage female participants to report positive feelings towards those with disabilities. Thus, the experimenter's expectancy may pose threat to the internal validity. As we know, the double-blind experiment is conducted to control the experimenter's expectancy. In double blind experiment both the participants as well as those in contact with the participants (experimenter's assistant) are in a sense, blind to the details of the experiment.

Like internal validity, the external validity of the research or experiment is threatened by a number of extraneous variables or factors. In general, all those factors which pose a threat to internal validity, also pose a threat to external validity. However, there are other additional factors which tend to reduce the power of the experiment to generalize the obtained results to a wider population of interest. Some of these factors, as discussed by Campbell and Stanley, are presented below.

1. *Artificial situation of the experiment*: The researcher tries his best to control all extraneous variables so that they may not produce any experimental change. As a consequence of this effort, the experimental situation becomes more artificial and less resembles the real life situation regarding which generalizations are to be made. The overall result is that on the basis of such an experimental situation, no concrete and sound generalization can be made.

2. *Interaction effects of selection bias*: We have seen how selection bias poses a threat to the internal validity of the research. It becomes a threat to the external validity when the researcher does not draw a random sample or a representative sample from the specified population. When the samples are not representative of the population, it becomes very difficult for the researcher to generalize the conclusion. Psychological, educational and sociological researchers rarely, if ever, utilize a random sample. Most of these researches are conducted on classroom samples, which are easily available to the researchers. Hence, the generalization from such studies can be made with a considerable risk. In order to eliminate the selection bias, as it relates to external validity, it is essential for the researcher that he must draw a random sample from the population.

3. *Prior knowledge about subjects*: When the researcher has some prior knowledge about the subjects, he is likely to show either personal bias in form of favour, or he may provide some clue that will affect the objective judgement of the subjects. In either case, sound generalization from the experimental setting becomes difficult.

4. *Carry-over effects of prior treatment*: When the experimental design is such that the same subjects serve as members of both the control and the experimental conditions, the chances are that the effect of the first treatment will not be fully wiped out and thus the effect is carried to the second treatment. The effect may be facilitating or interfering, depending upon the nature of the problem being studied. In either case, generalization becomes hazardous.

5. *Hawthorne effect*: This effect was first pointed out in 1920 by Mayo, Roethlisberger and Dickson in a series of experiments conducted at the Hawthorne plant of the Western Electrical Company in Chicago. The purpose of the experiments was to examine the effect of certain working conditions upon the workers' output. The intensity of light was one of the manipulated variables. The investigators found that as the intensity of light was increased, the workers' output also increased. After reaching a peak, the investigators gradually decreased the intensity of light and surprisingly enough, again the workers' output increased instead of decreasing. It was concluded that the workers' awareness of participation and the attention given by them were important motivating factors, which led to the increased output even during the poor level of illumination. From these studies, the term 'Hawthorne effect' was introduced and it refers to the increments in performance prompted by the subjects' awareness of participation in an experiment. When subjects demonstrate the Hawthorne effect in any experiment, the results of the experiment are likely to differ from the results of those experiments where no such effects are being demonstrated. Therefore, the conclusions thus reached become difficult to be generalized.

Since there are several perspectives from which one can generalize the obtained results in an experiment or study, there are two additional aspects of external validity which must be examined: *temporal validity* and *ecological validity*.

Temporal validity refers to the extent to which generalizations from a study are accurate across different time periods. It has two applications as follows:

(i) A treatment in a study may last for a particular period of time and there may also be a fixed period of time between a treatment and measurement of dependent variables. Here, temporal validity is the extent to which an observed relationship between the variables can be generalized to other time frames.

(ii) A study occurs in a specific time, that is, in a particular month or year. Here temporal validity is the extent to which results can be generalized to other months or years.

Many studies done in the past are today still considered valid. Studies of basic psychological processes, namely, thinking, memory and emotion, motivation are still valid today. However, studies of various social processes like attitude, opinion and other such behaviours which are influenced by societal changes are less well generalized today. Therefore, study of such social processes have poor degree of temporal validity.

Ecological validity is the extent to which the situation and behaviours in a study are like those found in natural settings or environment. In other words, ecological validity is the extent to which the experimental situation can be generalized to natural settings. For example, many experiments done in the field of verbal learning and memory utilize nonsense syllables or consonant syllables. But such syllables are not an everyday experience for most people. Therefore, such a learning process is an unusual one that occurs only in the laboratory. Thus such research lacks ecological validity. We can never be confident that it accurately generalizes to and describes natural and day-to-day learning processes of the person. In this way we can finally conclude that if the research design does not have ecological validity, we end up focusing on what the subjects *can* do in the experiment and not what they *usually* do in their lives.

CONTROLLING THREATS TO RELIABILITY AND VALIDITY IN RESEARCH

Threats to reliability and validity of a research are shown due to a variety of conditional factors. Some of them have already been discussed. Here, we shall focus on those ways that try to anticipate and eliminate anything that may threaten the reliability and validity. Following are some such important ways:

(i) *Doing a reliable manipulation*: Reliable manipulation of independent variables is one way to arrest such threats. As we know, reliability means that our results are consistent and contain no error. Reliability is sought not only in terms of measuring subjects' scores but is also sought in terms of manipulating the independent variable. By reliable manipulation of independent variable is meant that all subjects in one condition must receive the same treatment, that is, the same amount of independent variable, and when the experimenter changes to another condition, all subjects there should receive the same, new amount of independent variables. For example, if the experimenter is interested in knowing the impact of degree of temperature upon work performance, and if he selects two levels of conditions of temperature say, 60 degree F and 90 degree F then he wants precisely 60 degrees for all subjects in one condition and 90 degrees for all subjects in the other condition.

If there is inconsistency in our manipulations, the conclusion will be biased on the wrong amounts of the independent variables. Apart from this, if the experimental situation is in any way different for each subject (due to inconsistency in manipulation of independent variables), we shall get different responses and thus, variability among scores within each condition will go up producing higher degree of error variance. The greater the error variance, the poorer the strength of relationship and dependability upon the variables under study.

(ii) *Control of confounding variables:* In any experiment, the experimenter seeks to draw a causal inference which is of paramount importance for internal validity. The experimenter must be confident that only changes in independent variables are the real causes for changes in dependent variables. For this it is essential that the manipulation of independent variable should not introduce *confounding* which is said to have occurred when some extraneous variables are left uncontrolled, and therefore, systematic differences between the conditions exist, in addition to the independent variable. The process of confounding can be explained through an example. Suppose the experiment is being conducted in order to know the impact of room temperature on the aggressiveness of the subjects. Now further suppose that this experiment has been designed in such a way that in one condition a male experimenter is present when the room has a temperature of 70 degrees, and in the other condition a female experimenter is present when the room has a temperature of 90 degrees. In this case, we cannot tell with certainty whether differences in subjects' aggressiveness are due to differences in room temperature or to differences in the sex of the experimenter. This is because of the fact that variable of temperature confounds with the sex variable.

It is essential that the experimenter must employ a design that eliminates potential confounding. There are several ways to eliminate potential confounding by controlling extraneous variables. The process of elimination, balancing and randomization can be successfully employed for such elimination and control. For other such methods, the reader should refer to Chapter 20.

(iii) *Defence against diffusion of treatment:* Diffusion of treatment also threatens internal validity associated with the independent variable. In fact, diffusion of treatment occurs when subjects in one condition become somehow aware of the treatment given in other conditions. Generally it happens in between-subjects or between-group design in which subjects who have participated in one condition tell other potential subjects of other group/condition about their treatment / experience. As a consequence, these potential subjects are now aware of the treatment to be given, and hence the strength or impact of the manipulation is reduced. Not only this, the internal validity is also threatened, that is, subjects, instead of being influenced by the condition of the independent variable, are now influenced by the extraneous information about the condition of the experiment.

It is, therefore, essential that some defensive steps be taken against such diffusion. These are—

(a) One line of defense is to explain to the subjects why it is relevant and important that they should not tell other potential subjects about the details of the experiment.

(b) The whole experiment or study should be conducted in a very brief period of time so that the subjects may not get time to communicate with each other.

(c) Subjects might be tested from different locations so that they have little physical contact with each other and thus, little opportunity to communicate with each other.

(d) If possible, the treatments can also be disguised.

(iv) *Adoption of consistent procedures:* A very important issue in designing the procedures of the study is the need to produce consistency within and between conditions. For achieving this consistency, a careful planning is needed at the following front:

(a) Instructions: The instruction to be given in the study/experiment should clearly explain the task so that subjects may consistently attend to the stimulus and respond appropriately. Particularly, the experimenter should describe the sequence of events, identify the stimuli they should attend to and explain how to indicate a response. All these will create consistency, that is, all subjects would perform precisely the same intended task, without introducing extraneous stimuli/behaviour that make the task different for different subjects.

(b) Automation: Automation provides better consistency with respect to the presentation of stimulus. For example, when a stopwatch is used for controlling time for presenting visual stimuli, it may introduce inconsistency because there may occur variation in actual presentation time. Therefore, the best alternative is to rely on automated equipment-electronic timers, slide projectors, and the like for controlling and presenting the stimuli.

(c) Pilot studies: A Pilot study is defined as the miniature version of a final study in which the experimenter/researcher tests the procedures of the final study. In pilot study a sample of the type of subjects that will be used in final study is tested and as a consequence, the experimenter determines such matters as whether the given instructions are clear and the subjects perform accordingly, whether the task is doable within the time constraints or other demands, and whether the manipulation actually works and is strong. In addition to all these, the pilot study also allows the researcher to detect any bugs in the equipment or procedure. In this way, a pilot study provides the experimenter a good background for providing consistency of procedure.

(d) Group testing: Testing subjects in a group is more common and the advantage of group testing is greater efficiency in collecting data. Apart from this, in group testing all subjects are tested in the same condition at a time, and thus all subjects experience the same condition consistently presented. The disadvantage is that the subjects may create unnecessary noise which may provide distraction to some other subjects. Therefore, such extraneous behaviours have to be controlled so that the potential confoundings may be prevented.

(e) Manipulation checks: Manipulation check is a sort of measurement which determines whether the manipulation of independent variable has influenced the behaviour of the subjects in the environment. Usually such a check is made after the subjects have performed the task. It is particularly useful for ensuring that an independent variable has the intended influence on the intervening variable. For example, if the researcher manipulates frustration to influence the subject's intervening anger, and subsequently, measure their aggressiveness, he can check the manipulation by directly measuring the subject's anger. If the subjects make the desired responses, he can put greater confidence in concluding that differences in aggression are the result of differences in anger. If such a thing does not happen, the experimenter may go backward for redesigning the study.

In this way, we find that there are several ways through which threats to reliability and validity in an experiment can be controlled.

PHASES OR STAGES IN RESEARCH

The purpose of this section is to provide the potential researcher with important steps needed for conducting a scientific research. An experienced researcher knows that research is a tedious, painful and slow-moving job. However, if he follows certain steps in conducting the research, the work can be carried smoothly with least difficulty.

Identifying the Problem

The first step in conducting a research is to identify the problem. The researcher must discover a suitable problem and define it operationally. A problem is defined as that interrogative statement, which shows a relationship between two or more variables in an unambiguous manner. For example, take the following statement:

What is the relationship between academic ability and socio-economic status?

A problem has several other characteristics which become the relevant considerations in choosing a scientific problem. For identifying a good solvable problem, the investigator undertakes the review of the literature. A body of prior work related to a research problem is referred to as the *literature*. Scientific research includes a review of the relevant literature. When a researcher reviews the previous researches in related fields, he becomes familiar with several knowns and unknowns. Therefore, one obvious advantage of a review of the literature is that it helps to eliminate duplication of what has already been done and provides fertile guidance and suggestions for further research. The main purpose of a review of the literature is fourfold. *First*, it gives an idea about the variables which have been found to be conceptually and practically important or unimportant in the related field. Thus, the review of the literature helps in discovering and selecting variables relevant for the given study. *Second*, the review of the literature provides an estimate of the previous work done. This has a twofold advantage—it avoids unnecessary duplication of previous work and provides an opportunity for the meaningful extension of the previous work. As we know, many researches in social science are concluded with a suggestion for further exploration. Only after the review of the literature can this fact be known. *Third*, a review of the literature helps the researcher in synthesizing the expanding and growing body of knowledge. This facilitates in drawing useful conclusions regarding the variables under study and provides a meaningful way for their subsequent applications. *Fourth*, a review of the literature also helps in redefining the variables and determining the meanings and relationships among them so that the researcher can build up a case as well as a context for further investigation that has merit and applicability.

There are different sources of review of the literature. Journals, books, abstracts, indexes, periodicals, etc., are the major sources of review of the literature. Chapter 19 has been fully devoted to the review of the literature.

Formulating a Hypothesis

When the researcher has identified the problem and reviewed the relevant literature, he formulates a hypothesis which is a kind of *suggested answer* to the problem. Hypothesis may be defined as a tentative statement showing a relationship between variables under study. It is stated in the form of a declarative sentence. For example:

1. Persons coming from the upper and middle socio-economic status have stronger academic ability than persons coming from a lower socio-economic status.
2. Reward facilitates learning.

A good research hypothesis has several characteristics. *First*, the hypothesis should be consistent with the known facts and theories. *Second*, it should be testable. In other words, the hypothesis should be such which can be reinstated in a way that the researcher can test it and show that it is probably *true* or probably false. *Third*, a hypothesis should be reasonable and expressed in the simplest possible words.

For unbiased research, the researcher must formulate a hypothesis in advance of the data-gathering process. No hypothesis should be formulated after the data are collected. Chapter 18 presents a detailed discussion regarding the different aspects of hypothesis.

Identifying, Manipulating and Controlling Variables

Variables are defined as those characteristics which are manipulated, controlled and observed by the experimenter. At least three types of variables must be recognized at the outset—the dependent variable, the independent variable and the extraneous variable. *The dependent variable* is one about which the prediction is made on the basis of the experiment. In other words, the dependent variable is the characteristic or condition that changes as the experimenter changes the independent variables. The *independent variable* is that condition or characteristic which is manipulated or selected by the experimenter in order to find out its relationship to some observed phenomena. An *extraneous variable* is the uncontrolled variable that may affect the dependent variable. The experimenter is not interested in the changes produced due to the extraneous variable and hence, he tries to control it as far as practicable. The extraneous variable is also known as the *relevant variable*. In Chapter 20, a detailed discussion of the different types of variables commonly used in psychological, sociological and educational research is presented.

Formulating a Research Design

A research design may be regarded as the blueprint of those procedures which are adopted by the researcher for testing the relationship between the dependent variable and the independent variable. There are several kinds of experimental designs and the selection of any one is based upon the purpose of the research, types of variables to be controlled and manipulated as well as upon the conditions under which the experiment is to be conducted. The main purpose of the experimental design is to help the researcher in manipulating the independent variables freely and to provide maximum control of the extraneous variables so that it may be said with all certainty that the experimental change is due to only the maniplation of the experimental variable. In Chapter 21, a detailed discussion of the different types of research designs has been presented.

Constructing Devices for Observation and Measurement

When the research design has been formulated, the next step is to construct or collect the tools of research for scientific observation and measurement. Questionnaires, opinionnaires and interviews are the most common tools which have been developed for psychological, sociological and educational research. All these tools of research are ways through which data are collected by asking for information from persons rather than observing them. These techniques have been discussed separately in Chapter 12.

Summarizing Results

The next step in scientific research is to summarize the results so that a suitable analysis can be made. There are two common methods for summarizing results—the tabular method and the graphic method. In the tabular method the obtained data are reduced to some convenient tables, which facilitate the use of appropriate statistical tests. In the graphic method, the obtained data are shown through graphs and pictures. In general, the graphic method has an advantage over the tabular method in the sense that it provides quick deliberation and understanding to those who examine it. But the general limitation of the graphic method is that complex data are difficult to be displayed whereas the same can be easily shown through the tabular method. A good researcher, therefore, should summarize the results by utilizing both the methods.

Carrying out Statistical Analysis

When the data have been reduced to the tabular form, the next step is to carry out appropriate statistical analysis. There are two types of statistical tests—the parametric test and the non parametric test. Depending upon the nature of data and purpose of the experiment, either a parametric statistic or a nonparametric statistic is chosen for statistical analysis. In general, the purpose of carrying out the statistical analysis is to reject the null hypothesis so that the alternative hypothesis may be accepted. Commonly, there are two levels of significance at which the null hypothesis is rejected—0.05 level (or 5% level) and 0.01 level (or 1% level). These levels are also known as the alpha level. A separate chapter (Chapter 23) has been devoted to statistical analysis.

Drawing Conclusions

The investigator, after analyzing the result, draws some conclusions. In fact, the investigator wants to make some statement about the research problem which he could not make without conducting his research. Whatever conclusion is arrived at, he generalizes it to the whole population. At this stage, the investigator also makes some predictions about certain related events or behaviour in new situations.

TYPES OF EDUCATIONAL RESEARCH

It is very difficult to classify the types of educational research. This fact is further complicated due to the observation that almost every textbook suggests a different system of classification. However, the attempt made by Best & Kahn (1998) in this direction is a very scientific one and their classification is so wide and comprehensive that all researches fall under one of the following three types or a combination thereof.

1. *Historical research*: A historical research is one which investigates, records, analyzes and interprets the events of the past for the purpose of discovering sound generalizations that are helpful and useful in understanding the past and the present and, to a limited extent, the anticipated future. Thus, historical research describes *what was*.
2. *Descriptive research*: A descriptive research is one which describes, records, analyzes and interprets the conditions that exist. In such a research an attempt is made to discover relationship between existing nonmanipulated variables, apart from some comparison or contrast among those variables. Thus descriptive research basically describes what is. It is also known as *nonexperimental* or *correlational research*.
3. *Experimental research*: An experimental research is one in which the primary focus is upon the variable relationship. Here certain variables are controlled or manipulated and thier effect is examined upon some other variables. Thus experimental research basically describes what will happen when certain variables are carefully controlled or manipulated.

We shall discuss in detail the descriptive or nonexperimental research and experimental research in this chapter, whereas historical research will be discussed in Chapter 17.

TYPES OF RESEARCH: EXPERIMENTAL AND NONEXPERIMENTAL

Any attempt to classify behavioural research presents a difficult problem because there is no universally acceptable classification of research. Almost every textbook suggests a different system of classification of research and this fact in itself is a convincing evidence for the nonavailability of a universal system of classification. Despite these differences, an attempt is made here to present a near-common classification of research.

Based on the *application* of research study, the most general way of classifying research is to divide it into *fundamental* or *pure* or *basic research* and *applied research*. A fundamental research is the formal and systematic process where the researcher's aim is to develop a theory or a model by identifying all the important variables in a situation and by discovering broad generalizations and principles about those variables. It utilizes a careful sample so that its conclusion can be generalized beyond the immediate situation. However, a fundamental research has little concern with the actual application of its principles or generalizations. Applied research, as its name implies, applies the theory or model developed through fundamental research to the actual solution of the problems. It has many characteristics common to fundamental research. It also tends to make generalizations about the population and uses the various sampling techniques in selecting samples for its study. The main purpose of applied research is not to develop theories about a fact but to test those theories in actual situations. One popular type of applied research is *'action research'*. In action research the researcher emphasizes a problem which is immediate, urgent and has local applicability. Thus, the researcher here focuses upon the immediate consequences and applications of a problem and not upon general or universal application nor upon the development of a theory or a model. A teacher may undertake a research to know the reasons underlying unhealthy classroom habits so that the immediate outcome may benefit the local classroom students. Types of applied research will be discussed later in this chapter.

Depending upon the *objective* of the research, it may be classified into *experimental research* and *nonexperimental research*. An experimental research is one where the independent variables can be directly manipulated by the experimenter, and participants or subjects are randomly assigned into different treatment conditions. In the social sciences, experimental research is sometimes called *S/R research* because the researcher manipulates a stimulus (or stimuli) in order to establish whether or not this produces a change in a certain response (or responses). It is further divided into two main types—laboratory experiments and field experiments. A nonexperimental research is one where independent variables cannot be manipulated and therefore, cannot be experimentally studied. The subjects, too, are not randomly assigned into different treatment conditions. In nonexperimental research, also called post facto research, the responses of a group of subjects are measured on one variable and then compared with their measured responses on another variable. Such research is also called *R/R research* because changes in one set of responses (*R*) are compared with possible changes in another set of responses (*R*). A nonexperimental research, or descriptive research, can be divided into six main types—*field studies, ex post facto research, survey research, content analysis, case study* and *ethnographic study*. Perhaps more than half of the researches in psychology, sociology and education are nonexperimental. A detailed discussion of each of these six types of researches as well as the two types of experimental research is discussed ahead.

Based on *inquiry* mode, there are two types of research—*quantitative research* and *qualitative research*. A quantitative research is a research where objectives, design, sample and statistical analysis are predetermined. It is also sometimes known as *structured research*. A qualitative research is one which allows flexibility in all these aspects of the process. A detailed discussion of qualitative research has been done in chapter 22. This research is also known as *unstructured research*. In general, a quantitative research is more appropriate for determining the *extent* of the problem whereas qualitative research is more appropriate for exploring the *nature* of the problem. Figure 15.1 presents the classification of research from these three perspectives (application, objectives and inquiry).

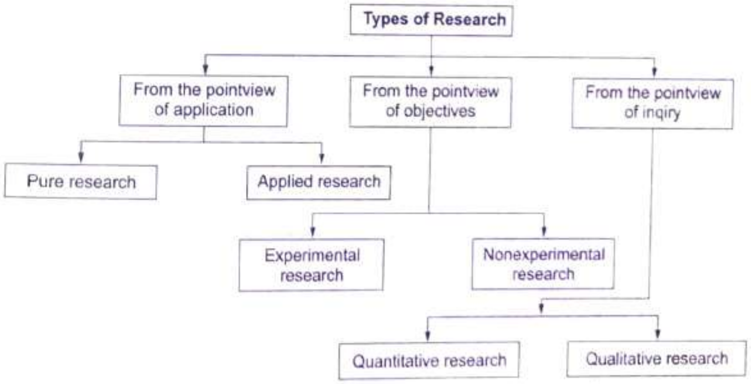

Figure 15.1 Classification of research from different modes

Laboratory Experiments

A laboratory experiment is one of the most powerful techniques for studying the relationship between variables under controlled conditions. It may be defined as the study of a problem in a situation in which some variables are manipulated and some are controlled in order to have an effect upon the dependent variable. The variables which are manipulated are known as *independent variables* and the variables which are controlled, are known as *extraneous* or *relevant variables*. Thus in a laboratory experiment, the effect of manipulation of an independent variable upon the dependent variable is observed under controlled conditions. Festinger & Katz (1953, 137) have defined a laboratory experiment as "one in which the investigator creates a situation with the exact conditions he wants to have and in which he controls some, and manipulates other variables." Analyzing this definition, two salient features of a laboratory experiment emerge.

First, in a laboratory experiment, the experimenter creates a situation in which all the possible extraneous variables are controlled so that the variances produced by them are also controlled or kept at a minimum.

Second, in a laboratory experiment the variables are manipulated (called independent variables) and the effect of manipulation of these variables upon the dependent variable is examined.

There are several ways of manipulating and controlling the variables, which have been discussed separately in Chapter 20. However, one of the most common methods used in behavioural research is the use of pre-experimental instruction. At the start of the experiment, the subjects are given some instructions verbally in a manner that the variables are automatically manipulated. For example, if one wants to examine the effect of competition upon the performance of a task, one may take two groups of subjects matched in age, sex and intelligence. Subsequently, one may impart instruction to one group in a way which enhances the feeling of competition and impart a neutral instruction to the other group. For instance, for enhancing the sense of competition one may say that all subjects would be ranked in terms of their individual performance and that the person who contributes most would get the highest rank and the person who contributes the least would get the lowest rank. In a group where no sense of competition is to be introduced, instruction may simply specify that the members of the group are required to perform the task as accurately and quickly as possible. In the above experiment, the competition was the independent variable and the performance of the task was the dependent variable. One group was instructed in a way to enhance the competition (the independent variable) and another group was a neutral group having no instruction for enhancing the feeling of competition. The former group is known as the *experimental group* and the latter group is known as the *control group.*

Following Kerlinger (1986), there are three main purposes of a laboratory experiment. *First,* a laboratory experiment purports to discover a relationship between the dependent variable and the independent variable under pure, uncontaminated and controlled conditions. When a particular relationship is discovered, the experimenter is better able to predict the dependent variable. *Second,* a laboratory experiment helps in testing the accuracy of predictions derived from theses or researches. *Third,* a laboratory experiment helps in building the theoretical systems by refining theories and hypotheses and thus, provides a breeding ground for scientific evaluation of those theories and hypotheses.

A laboratory experiment has some strengths and weaknesses. Its primary strengths or advantages may be enumerated as mentioned below.

1. A laboratory experimenter studies the problem in a pure and controlled situation. In fact, this is one of the greatest merits of a laboratory experiment. He sets up a condition where extraneous variables are maximally controlled so that the dependent variable is influenced only by the manipulation of the independent variables. The conclusion drawn from a problem being studied in such a controlled condition is more dependable and thus the laboratory experiment has the fundamental requisite for any investigation, that is, internal validity.

2. A laboratory experiment is replicable. If one has some doubt over the conclusion reached by a particular experimenter, he can replicate the design, conduct the experiment and verify the conclusion. A laboratory experiment is also replicated when one wants to substantiate or refute the findings of earlier laboratory experimenters.

3. A laboratory experiment provides a great degree of precision in manipulation of the independent variables. Not only this, if the laboratory experimenters want to assign the subjects randomly to different treatment conditions, they can do so easily with minimum risk of introducing subjectivity in the controlled situation.

4. A laboratory experiment has sufficient degree of internal validity because the experimenter usually has the maximum possible control over the extraneous variables and the manipulation of independent variables.

Despite its strength, a laboratory experiment has some weaknesses, too, as indicated below.

1. A laboratory experiment lacks external validity, although it has a sufficient degree of internal validity. When a laboratory experimenter, for example, reaches some conclusions regarding the rate of learning a maze by a rat, can he generalize it to the human maze learning? The answer is, yes, to some extent. Even when he does so, he does this at considerable risk. This lack of external validity makes the laboratory experimenter unable to make generalizations with full confidence.

2. The experimental situation of a laboratory experiment is said to be an artificial and a weak one. But this criticism appears to be inappropriate because it results from inaccurate understanding of the purpose of a laboratory experiment. In fact, a laboratory experiment should not, and need not, be an exact duplication of a real-life situation. If one wants to study some problems in a real-life situation, he should not bother to set up a laboratory experiment duplicating such a situation. A laboratory experiment simply creates a situation where the variables are controlled and manipulated under specially defined conditions. Such a situation may or may not be encountered in real life. In fact, some laboratory situations are such that they can never be encountered in real life. It may, therefore, be concluded that the criticism regarding the creation of an artificial situation comes mainly from those persons who misunderstand the purpose of a laboratory experiment.

3. In most laboratory experiments, a complex design is followed and as a consequence, more than two variables are simultaneously manipulated. It is a matter of common observation that if more variables are manipulated simultaneously, the strength of each variable is lowered. This is particularly obvious in those laboratory situations where the manipulation of the variables is done through verbal instructions.

4. According to Robinson (1976), there are some situations like mass riots which can't be studied in a laboratory because of sheer physical impossibility.

5. For want of social approval, the behaviour of persons in certain situations may not be studied by laboratory research. For example, if the experimenter is interested in studying the impact of deprivation of visual sense modality upon depth perception in children, he may not be able to carry out the laboratory experiment, because no parents will allow their children to experience such harmful deprivation.

6. Laboratory-experiment research is costly as well as time-taking. Some of the apparatuses are so costly that the experimenter normally fails to buy them and hence, remains unable to conduct the experiment. Not only this, some experiments particularly those relating to longitudinal genetic studies on humans, can't be successfully carried out because the time required to carry through even a few generations would be very long.

Despite these weaknesses, laboratory experiments are the greatest achievements of behavioural scientists. Such experiments possess the most important prerequisites for any behavioural investigation, that is, the internal validity, although they have the trait of poor generalizability.

Field Experiments

A field experiment is one of the common research tools in the hands of sociologists, educationists and social psychologists. It is very similar to a laboratory experiment. A field experiment may be defined as a controlled study carried out in a more or less realistic situation or field where the experimenter successfully manipulates one or more independent variables under the maximum possible controlled conditions. Following Shaughnessy & Zechmeister (1990), when the experimenter manipulates one or more independent variables in natural setting for determining their effect upon behaviour, the procedure is known as field experiment. In sum, a field experiment is an experiment conducted in natural setting. The setting may be a school, a factory, a hospital, a shopping complex, a street corner or any place in which the behaviour can be identified. Such experiments are common when the research is more applied to or seeks to examine complex behaviours that occur in natural setting. Analyzing this definition, we get three main features of a field experiment.

First, a field experiment is carried out in a more or less realistic situation. Thus, it differs from a laboratory experiment which is carried out in the artificial situation of a laboratory.

Second, the field experimenter, like the laboratory experimenter, also manipulates the independent variables and controls the extraneous variables.

Third, a field experimenter manipulates the variables under as carefully controlled conditions as the situation permits. On this point, a field experiment differs from a laboratory experiment because in the latter case the situation is controlled in all possible respects whereas in the former case, the exercise of control is dependent to the extent it is permitted by the situation. As Kerlinger (1973, 401) has opined, "Sometimes it is hard to label a particular study 'laboratory experiment' or 'field experiment.' Where the laboratory experiment has a maximum of control, most field experiments must operate with less control, a factor that is often a severe handicap to the experiment."

Like the laboratory experiment, the field experiment also has strengths and weaknesses. The principal strengths may be enlisted as follows:

1. A field experiment deals with the realistic life situation. Hence, it is more suited for studying social changes, social processes and social influences. Several field experiments have been conducted for studying these social processes and fruitful results have been obtained.

2. Since in a field experiment observations are done in natural settings, such studies generally have a high degree of external and ecological validity. The obtained findings are generalized to the real world because they are obtained in the real world.

3. In a field experiment, we can observe behaviour when subjects are psychologically engaged in a real situation. As a consequence, we have greater experimental realism which is defined as the extent to which the experimental task engages subjects psychologically such that they become less concerned with demand characteristics. Such advantage is especially important because instead of depending upon potentially unreliable self-reports of what subjects say they do, we can observe what they are really doing in the situation.

4. A field experiment allows us to go into the real world and replicate the laboratory studies to insure the generality of their findings. After knowing about a behaviour in the laboratory, we go out and see whether it operates the way we think it does.

The principal weaknesses of field experiments are as given below.

1. Since a field experiment is carried out in a realistic situation, there is always the possibility that the effect of independent variables is contaminated with uncontrolled environmental variables. The unexpected noise and gatherings may affect the dependent variable and thereby, contaminate the influence of the independent variable. In a laboratory experiment, this problem does not arise because of the fully controlled laboratory situation. However, if the situation is somehow fully controlled in a field experiment, it would prove to be a more powerful tool than a laboratory experiment.

2. In many field situations, the manipulation of independent variables may be difficult due to non-cooperation of subjects. For example, suppose the investigator wants to test whether or not frustration leads to aggression in a group of small children. He may be producing frustration in one group of children (called the experimental group) and may be keeping another group of children, a neutral one, that is, free from frustration. If the parents of the children come to know that their wards are to be exposed to frustrating situations, they may not like it and may restrain their children from being exposed to field situations. Likewise, numerous examples of a field situation can be cited where the manipulation of independent variables may pose difficulties to the field experimenter due to similar reasons.

3. In a field experiment, it is not possible to achieve a high degree of precision or accuracy because of some uncontrolled environmental variables. But in the case of a laboratory experiment, it is possible to achieve a high degree of precision because all extraneous variables are controlled to the maximum possible extent.

4. A field experiment requires that the investigator has high social skills to deal effectively with people in a field situation. It automatically implies, then, that an investigator with poor social skills will not be very effective in conducting a field experiment. It has also been observed that even where the investigator possesses a high degree of skill, a field experiment usually takes such a long time that subjects' cooperation with a uniform spirit becomes doubtful. This tends to lower the validity of the experiment.

Field Studies

Any ex post facto scientific study which systematically discovers relations and interactions among variables in real life situations such as a school, factory, community, college, etc., may be termed as a *field study.* There are two important features of a field study.

First, a field study is an ex post facto study and an ex post facto study is one where the investigator tries to trace an effect that has already been produced to its probable causes.

Second, in any field study no independent variables are manipulated and thus it differs from a field experiment where the independent variables are manipulated for determining relations among variables. In a field study, the investigator depends upon the existing conditions of a field situation as well as upon the selection of subjects for determining the relationship among variables.

Although field studies bear some similarity to survey studies, the two types of researches differ. *First,* in a survey the emphasis is upon the selection of a representative sample to make an accurate description of the characteristics of the larger universe (survey research) and the characteristics of a fraction of the universe (sample survey). But in field studies the emphasis is upon the processes under investigation, rather than upon their typicality in the universe or population (Katz 1953). The investigator in a field study, therefore, may or may not give that emphasis to representative sampling. *Second,* the investigator in a field study generally deals with a single group, community or section for analyzing the social and psychological processes whereas in survey studies, the investigator generally takes a cross-sectional group or cross-community for analyzing social and psychological processes. Thus, in a field study the group being studied is small as compared to survey studies and therefore, it can be reasonably expected that a field study will provide a better picture of social interactions or interrelations than survey studies.

Types of Field Studies

Katz (1953) has divided field studies into two types:

(a) Exploratory field studies

(b) Hypothesis-testing field studies

A discussion of these two follows.

(a) *Exploratory field studies*: Exploratory field study is one that intends to discover significant variables in the field situation and find out relations among those variables so that the groundwork for better and more systematic testing of hypotheses can be laid obviously. Thus, exploratory field study seeks *what is*; it does not rather seek to *predict relations* to be found later. On the basis of the results of such field study, the investigator is able to find out a relationship between the variables, but he remains unable to provide concrete proof of the existing relationship. Suppose the investigator wants to investigate the correlates of productivity in a factory through exploratory field study. He may correlate productivity with several factors like age and sex of the employees, period of work, pay, etc. But he will not be formulating any hypothesis relating to the productivity and any of these variables beforehand. It may be that on the basis of such field studies, he may provide some facts which would be explored in future.

(b) *Hypothesis-testing field studies*: Hypothesis-testing field study is one in which the investigator formulates some hypotheses and then proceeds to test them. He provides some concrete evidence for such testing. Thus, here the investigator aims at predicting relations among variables. Suppose, the above study is done by hypothesis-testing field study, the investigator will proceed by formulating certain hypothesis and subsequently, he will test its reliability and validity. He may, for example, formulate the hypothesis that poor pay for prolonged hours of work may result in lower productivity. On the basis of the results obtained by hypothesis-testing field study, he will verify the truth of the hypothesis.

Both these types of field studies are common but the hypothesis-testing field study is more popular than exploratory field study.

The principal advantages of field studies are give below.

1. A field study provides opportunities for direct observation of social interaction and relationships. The investigator can directly make an observation of how people actually behave when placed in a group situation.

2. A field study is usually carried out in a realistic situation like a school, college, factory, community, etc. As such, it avoids the artificialities of laboratory experiments. A field study may, therefore, be more useful for an educator and a sociologist rather than for a psychologist.

3. In a field study, a continued observation for a given period of time is possible because it tends to persist for that period of time. One advantage of this continued observation, says Katz (1953), is that "the timing of certain variables may be ascertained."

4. The investigator in a field study is allowed, as Katz says, to record "reciprocal perceptions and interdependent reactions" from groups of people. One advantage of the reciprocal perceptions and interdependent reaction is that they give a total picture of the social structure, the complexity of which might otherwise be missed.

Despite these advantages, a field study has the following weaknesses:

1. Generally, field situations where the field studies are carried, are so complex that they make the precise measurement of the variables a very difficult task. When variables cannot be measured precisely, it adversely affects the internal validity and subsequently, the external validity of the studies.

2. In any field study, there are a large number of variables. In experimental research like a laboratory experiment or field experiment, these variables can be controlled with greater satisfaction. But in field studies, these variables cannot be fully controlled or can be controlled with less satisfaction. Again, this tends to lower the internal validity of a field study.

3. A field study also suffers from lack of practicality. It generally takes a longer time; its cost is high; and its samples are usually large. All these factors force the investigator to avoid a field study as far as possible.

Difference between Field Experiment and Field Study

Field experiment and field study are two important modes of social scientific research. Both are similar in the sense that both these research studies are conducted in fields, that is, in a realistic situation. Another similarity is that both avoid the use of controlled conditions of the laboratory. Despite these similarities, a field experiment is different from a field study as indicated below.

1. A field experiment is an experimental research whereas a field study is a nonexperimental research.

2. In a field experiment, the independent variables are manipulated and its impact upon dependent variables is examined. In a field study the investigator does not manipulate variables; rather he aims at discovering the relations and interactions among psychological, sociological and educational variables.

3. A field experiment is more precise than a field study. In a field experiment the experimenter is able to maintain control to a greater extent and, therefore, precision in measurement of field variables is maintained. But in field studies, the problem of such precision is more acute.

Ex Post Facto Research

An ex post facto research is one in which the investigators attempt to trace an effect which has already occurred to its probable causes. Thus the term ex post facto research means that the researcher has conducted the study after the events have occurred. (The phrase ex post facto means "after the fact".) The effect becomes the dependent variable and the probable causes become the independent variable. Thus in the ex post facto research the manifestation of independent variables occurs first and then its effect becomes obvious to the investigator. Since the independent variables have already occurred, the investigator has no direct control over such variables. As such, the purposeful manipulation of the independent variable becomes difficult. A simple definition of ex post facto research may be formulated as given below.

> Ex post facto research is that empirical investigation in which the investigator draws the inference regarding the relationship between variables on the basis of such independent variables whose manifestations have already occurred. In this type of research, the investigator has no direct control over the independent variables because they occur much prior to producing their effects.

Suppose the investigator takes a case of lung cancer and then goes back to explore the probable causes of it. He may find that cigarette smoking, and a chronic cough are most commonly associated with lung cancer. Here, lung cancer is the dependent variable and cigarette smoking and the chronic cough are examples of independent variables. These two independent variables have already occurred and therefore, are beyond the direct control of the investigator. It is difficult for the investigator to control these independent variables either by manipulation or by randomization. Take another example. Suppose, the investigator wants to study the major determinants of academic achievement among primary-school-going children. After a review of the literature, he may find that three factors are most likely to produce differences in academic achievement. The three factors are SES (socio-economic status), motivation and intelligence. After treatment of the data he may find that there is a significant difference in academic achievement of upper SES children, middle SES children and lower SES children. The first group and the second group of children achieve higher than the third group. He, then, concludes that SES is one of the determinants of academic achievement. Likewise, he may find that higher motivation and higher intelligence are associated with higher academic achievement. In the above example, the nature of an ex post facto study is obvious. The dependent variable is the academic achievement and the independent variables are SES, motivation and intelligence, over which the investigator has no direct control.

Ex post facto research is also known as *causal-comparative research* or when correlational analyses are to be made, it is referred to as correlational research (Best & Kahn 1998).

The nature of ex post facto research can be made clearer by distinguishing it from experimental research. The following points may be enumerated as the points of distinction between these two types of researches:

1. In experimental research, the investigator has direct control over the independent variables. By direct control we mean the investigator's power to manipulate the independent variables. When one wants to study the effect of reward upon learning how to spell words in children, one may give a reward to one group of children, and have a control group of matched children who are not given a reward for learning the spellings of words. Here the investigator is directly manipulating the independent variable, that is, the reward. In ex post facto research, the investigator has no direct control over the independent variables probably because of two reasons. First, the independent variables occur prior to the effect they produce and second, some independent variables used in ex post facto research are inherently not directly or experimentally manipulable. Variables like home backgrounds, school environment, aptitude, intelligence, parental inference are not manipulable. Thus ex post facto research lacks the direct control of independent variables.

2. In experimental research the random assignment of subjects to different groups or assignment of measurements to different groups at random can be easily done but in the ex post facto research, random assignment of subjects is not possible because the investigator is bound to take things as they are.

3. In experimental research, the investigator can make a statement regarding the relationship between the dependent variable and the independent variables with more certainty than in ex post facto research. This is because in the former the investigator can manipulate the independent variables and hence, can better relate the changes in the independent variables with changes produced in the dependent variables. This is not possible in ex post facto research.

Ex post facto research has some strengths and weaknesses. Its principal strengths or advantages may be enumerated as follows:

1. Ex post facto research is considered to be very important in behavioural researches where many variables are either not manipulable or not amenable to experimental enquiry. Many sociological and educational variables belong to these categories. However, these variables work satisfactorily with the controlled enquiry of ex post facto research.

2. In some circumstances, particularly when one wants to investigate causes on the basis of the effect, ex post facto research is more useful than experimental research.

Its principal weaknesses are given below.

1. In ex post facto research, as has been discussed earlier, the investigator cannot manipulate the independent variables. When the independent variables cannot be manipulated, the forecast regarding the relationship between the independent variables and the dependent variables becomes dubious.

2. In ex post facto research, the investigator cannot exercise control over independent variables through randomization. He cannot assign the subjects to different groups at random nor can he assign the various treatments to the different groups at random.

3. The third limitation which partly stems from the first one is that the investigator may not be able to provide a plausible explanation for the relationship between the independent and the dependent variables. Generally, the investigator falls prey to what is called post-hoc fallacy, a conclusion which says that because two factors go together, one is the cause and the other is the effect. Because there seems to be a close relationship between smoking and cancer, one may conclude that smoking is the cause and cancer is the effect, whereas in reality glandular imbalance may be the actual cause of cancer. Such imbalance may induce a certain amount of tension. Since excessive smoking is a tension-releasing mechanism, such persons would tend to be heavy smokers. In such a situation the cancer can result from glandular imbalance, rather than from smoking which is merely a symptom.

Despite these limitations, ex post facto research is a popular method of research for educational and sociological problems. Such researches are also being conducted in the field of psychology.

Survey Research

Survey research is a new technique for social science research. Survey, as such, is quite an old technique and was largely developed in the eighteenth century. However, in the second part of the nineteenth century a systematic literature was made available by Booth who is regarded as the father of scientific social surveys (Moser & Kalton 1971). But survey research, as a special branch of social science research, is considered as a new technique developed in the twentieth century.

Survey research is based on the simple principle that if one wants to find out what people think about some topic, just ask them. In other words, survey is a structured set of questions or statements given to people in order to measure their attitude, beliefs, values or tendencies to it.

Survey research, mostly used by psychologists, sociologists and anthropologists, should be distinguished from sample survey, which is its close ally. The survey researcher is primarily interested in assessing the characteristics of the whole population. Thus, survey research may be defined as a technique whereby the researcher studies the whole population with respect to certain sociological and psychological variables. For example, if a researcher wants to study how many people of both sexes in India adopt contraceptive devices as a measure of birth control, this will constitute an example of survey research. But a survey researcher rarely takes pains to make an approach to each member of the population or universe probably because it requires a lot of time, money and patience. Thus he takes a convenient random sample, which is considered to be representative of the whole universe and subsequently, an inference regarding the entire population is drawn. When a researcher takes a sample from the population for studying the relative incidence, distribution and relationship of psychological and sociological variables, the survey is termed as a *sample survey*.

Survey research depends upon three important factors.

1. As survey research deals with the characteristics, attitudes and behaviours of individuals or a group of individuals called a sample, direct contact with those persons must be established by the survey researcher.

2. The success of survey research depends upon the willingness and the co-operativeness of the sample selected for the study. The people selected for the survey research must be willing to give the desired information. In case they are not willing and do not co-operate with the survey researcher, he should drop the plan in favour of some other technique.

3. Survey research requires that the researcher be a trained personnel. He must have manipulative skill and research insight. He must possess social intelligence so that he may deal with people effectively and be able to extract the desired information from them.

Types of Survey Research

Depending upon the ways of collecting data, survey research can be classified into different categories, namely, *personal interview, mail questionnaire, panel technique* and *telephone survey*. A detailed discussion of each of them is presented below.

Personal Interview

Personal interview, also known as the *survey interview*, is one in which a direct conversation between the interviewer and the respondent is held with a view to elicit some information from the latter. The situation of a personal interview is such that the interviewer neither tries to help the respondent nor to educate him. He is simply interested in eliciting information from the respondent where he (the respondent) is likely to be one of many from whom similar information is elicited. According to Cannell & Kahn (Lindzey & Aronson 1968), there are three necessary conditions for a successful personal interview. The first is *accessibility*. By accessibility is meant that the information required be such that the respondent be able to convey it to the interviewer. In other words, the content of the required information must be expressible without producing any embarrassment to the respondent. The second is *cognition* by which is meant the understanding on the part of the respondent of what is required of him and what types of information are required. The respondent must know and understand what types of information he is going to express and in what terms of reference he should express the required information. Sometimes, it has been found that the respondent goes off the point and where this occurs, it becomes the interviewer's job to teach the respondent his appropriate role so that he may return to the right track. The third is *motivation* of the respondent. The respondent must be motivated to give accurate answers because a highly distorted answer is no better than no answer at all. When the respondent is motivated, he would tend to co-operate with the interviewer. Where the respondent lacks motivation, the interviewer should try to build up those factors, which may increase his motivation.

The success of personal interview is dependent upon the satisfactory fulfilment of these three factors. Apart from this, the success of an interview is largely dependent upon the interviewer's personality. An interviewer cannot be regarded as merely a means of extracting information. His personal bias and attitude may affect the required information. It is, therefore, suggested that the interviewer must be a trained person so that he may be able to ask probing questions in an impartial way and also be able to exhibit a permissive attitude throughout the interview. A detailed discussion of the interview technique is done in Chapter 12.

Mail Questionnaire

Mail questionnaire (or survey) is one of the most common types of survey methods used in educational and sociological researches. As its name implies, the questionnaire consisting of several items designed to elicit the required information is prepared and mailed to the respondent with the request to return it after answering all the items. Thus the mail questionnaire appears to be a direct means for obtaining information from every respondent. There are several *advantages* of a mail questionnaire.

First, the mail questionnaire is less costly than an interview. The cost of mailing questionnaires in which postal expenditures take the place of interviewers' expenditures is generally less than the cost of interviewing the respondent.

Second, through the use of mail questionnaires a widely scattered population can be surveyed rapidly and with less expenditure. Thus, a mail questionnaire is a quick method of survey even in the case of a scattered population.

Third, a mail questionnaire easily avoids many problems associated with the use of the interviewers. There are many sources of interviewers' errors, particularly their personal bias and influence, which tend to lower the reliability and the validity of the survey. The use of a mail questionnaire automatically avoids these errors of the interviewers.

Fourth, the problems arising out of contacting the respondent are avoided in the mail questionnaire.

There are ten principal *disadvantages* of the mail questionnaire.

1. The main problem is the nonresponse from the respondent. Generally, the percentage of responses to a mail questionnaire is very poor, although Scott (1971) has mentioned in his reports of five mail surveys carried out by the Government that the percentage of response rate was in the region of 90. In a mail questionnaire it is not the loss of sample numbers (due to the nonresponse of the respondent) which is serious, rather the probability that the personality characteristics of nonrespondent differ from the personality characteristics of the respondents and any inference based upon the latter is likely to be automatically biased. There can be two suggestions for improving the response rates of mail surveys. *First*, no awkward and/or embarrassing question should be asked because even a single awkward question is likely to produce a high rate of nonresponse. *Second*, the mail questionnaire should be accompanied by a stamped self-addressed reply envelope because many respondents do not wish to pay the postal charge and for that reason, may not reply. Besides, a third suggestion may also be made for improving the response rate to the mail questionnaire. If the design and the purpose of the study permit, those persons who are less educated, less interested and belong to the lower socio-economic status should not be included in the sample because evidences indicate that such persons do not reply to any questionnaire mailed to them.

2. Married women, especially those having above-average numbers of children, have a higher rate of nonresponse.

3. The mail questionnaire is an ineffective technique of survey where the objective and purpose of the survey needs sufficient explanation for its complete understanding.

4. The mail questionnaire is an effective technique where the questions to be asked are simple and straightforward so that they may be understood with the help of printed instructions. Where the questions are difficult, complex and technical, the mail questionnaires become useless.

5. When it is desirable to probe the respondents deeply or to talk with them, the mail questionnaires serve little purpose.

6. Mail questionnaires are inflexible techniques of survey. In other words, in a mail questionnaire there is no way to check the validity of the answers, to clarify the vague answers and to know the reasons behind the unwillingness to answer a particular question.

7. In a mail questionnaire, the different answers cannot be treated as fully independent because the respondent usually reads the whole questionnaire before he starts answering the questions.

8. In a mail questionnaire, the surveyor has no means by which he can be sure that the right person has answered the questions. Sometimes, it has been noticed that the questionnaire is completed by a person other than the right person. According to Scott (1971), there are two types of situations in which this usually happens. The first situation is one in which the questionnaire contains many questions which do not apply to the respondent. In this situation the respondent falsely thinks that the questionnaire is not meant for him and therefore, it should be passed on to persons who are more appropriate for answering the questions. The second situation is one in which the respondent gives little importance to the questionnaire and therefore, thinks that it matters little who responds to it.

9. When spontaneous answers are needed, mail questionnaires are not considered to be appropriate.

10. In a mail questionnaire, there is no way to supplement the answers with variable background data. The researcher cannot observe the mode of expression of the respondents, their reaction to typical questions and their general attitude towards the survey. All these observational data are lacking in the case of a mail questionnaire.

Some of the disadvantages of a mail questionnaire are such that they can be removed if they are combined with interviewing. For example, the questionnaire can be mailed and collected by the interviewers or interviewers can personally deliver questionnaires to each respondent who can mail them after completing them. When the interviewers collect the questionnaires, many problems arising out of ambiguity of the questions or printed directions can be solved. Similarly, personal delivery by the interviewers may solve the problem of wrong or incomplete addresses of the respondents because an interviewer can afford more time and take more pains than a postman in locating a respondent's house or flat.

Panel Technique

Some survey techniques require successive interviews with the same sample. The panel technique is one of them where the re-interview design is used and the same sample is interviewed more than once. Where the purpose of the survey is wide and extensive, multiple interviews are taken with the same sample. But where the objective of the survey is less extensive and wider, two interviews are sufficient. The panel technique has two *advantages.*

First, the panel technique enables the investigator to know how the various factors bring changes through time in the attitudes of the sample being studied.

Second, when the same sample is interviewed twice (or more than twice), it becomes a more sensitive and an accurate measure of change than when two different samples from the same population are tested.

The panel technique has, however, two important *limitations.*

First, in the panel survey there occurs a loss of the sample being studied. The loss may occur due to any of the three factors, namely, death, refusal to be reinterviewed, and moving from one place to another. The loss of a certain proportion of the sample naturally increases the probability of serious bias in the study.

Second, re-interviewing sometimes tends to sensitize the sample to the extent that the individuals refuse to give the desired response and may act as if they are no longer representatives of the population to which they belong.

Telephone Survey

Telephone survey is another form of survey research. In this survey, the respondent is interviewed by the investigator on telephone. The investigator calls a respondent on telephone, asks questions and records answers. Investigator samples respondents from the lists or telephone directories. Recently, some computer-assisted technologies have been developed for making telephone survey more smooth. Two such popular techniques are: *Computer-Assisted Telephone Interviewing* (CATI) and *Interactive Voice Response* (IVR). In CATI, the interviewer sits in front of a computer and makes calls. The interviewer wearing a headset and microphone, reads the questions from a computer screen for a specific respondent who is called. The interviewer records the answers via keyboard. Once the answer has been entered, the computer shows the next question on the screen. CATI has the advantages of speeding up the interview and reduces interviewer's errors. It also eliminates the separate step of entering information into a computer and speeds up data collection. With IVR, a respondent listens to the questions and response option over the telephone and responses are recorded through touch-tone entry or voice recognition software. IVR has some advantages like rapid and automated data collection, few errors and high anonymity. Researchers have shown that IVR is relatively successful for short and simpler surveys but has drop-off rates for longer surveys or questionnaires (Tourangean et al. 2002). This type of survey has the advantage of being quick and speedy in collecting information about the respondents. But the technique has several disadvantages. When the investigator is not known to the respondents, they usually do not co-operate and answer only simple and straightforward questions. In a country like India, the telephone survey has one additional disadvantage: Only a limited section of the population has telephone facilities and therefore, the chance is that not all respondents can be contacted on telephone. In such a situation, the telephone survey automatically defeats its purpose.

Advantages and Disadvantages of Survey Research

Survey research is one of the popular methods of research in behavioural sciences. It has some advantages and disadvantages. Its major advantages are given below.

1. Survey research has wide scope. In other words, through survey research a great deal of information can be obtained by studying the larger population. Although conducting a survey research is more costly than conducting a field experiment or a laboratory experiment, still, in view of the quality and amount of information rendered by a survey research, it can be taken as an economical method of research.

2. Survey research is more accurate. As Kerlinger (1986, 387) has put it, "The accuracy of properly drawn samples is frequently surprising, even to experts in the field. A sample of 600 to 700 individuals or families can give a remarkably accurate portrait of a community, its values, attitudes and beliefs."

3. Survey research has been frequently used in almost all the social sciences. Hence, the method has interdisciplinary value. In fact, such researches provide raw materials for a vast increasing 'gross disciplinary research' (Campbell & Katona 1953).

4. Survey research is considered a very important and indispensable tool for studying social attitudes, beliefs, values, etc., with maximal accuracy at the economical rate.

Despite the advantages, survey research has some disadvantages also as mentioned below.

1. Survey research remains at the surface and it does not penetrate into the depth of the problem being investigated. There are other types of researches which are preferred to survey research because they make deeper exploration of relations.

2. Survey research has some practical problems. Such researches are time-consuming and demand a good amount of expenditure. This is true, especially where a large survey is to be conducted in which interviews and schedules require skill, time and money.

3. Although it is true that survey research is accurate, it is still subject to sampling errors. In survey research there is always the probability of one chance in a twenty or hundred that an error, more serious than minor fluctuations of chance, may occur and distort the validity of the result obtained.

4. It has also been observed that some technique of survey research, such as survey interview, makes the respondent too sensitive and, hence, such a technique leaves the respondent out of his own social context. This usually invalidates the results of survey research.

5. Survey research demands expertise, research knowledge and sophistication on the part of the researcher. In other words, the competent researcher must know the technique of sampling, questionnaire construction, interviewing, analysis of data and other technical know-how of the survey. Most of the survey researchers don't possess these qualities in the required amount. This invalidates the quality of survey research.

6. Another problem in survey research is concerned with response bias, the most common of which is called a *social desirability bias.* It has been observed that sometimes people respond to a survey question in such a way that reflects not how they truly feel or what they really believe but how they think they should respond. That is, they tend to create a positive image of themselves— one that is socially desirable.

7. Still another limitation of survey research is concerned with the contents of items contained in the survey. Questions may be vague ones which can produce misleading results.

8. A final problem with survey research is not methodological but an ethical one. As we know, on the basis of survey research sometimes decisions affecting the lives of the people are made and if the surveys are flawed in any way, people can be hurt.

Despite these disadvantages, survey research has been a popular option not only for a psychologist but also for a sociologist, an educationist, an economist and a political worker.

Steps in Conducting Survey Research

The basic steps involved in conducting survey research can be divided into following six steps:

Step 1: The first step in survey research that can appropriately be called as *planning phase*, is the phase where the researcher develops an instrument—a survey questionnaire or interview schedule, decides on the type of survey (mail, interview, telephone, etc.), decides on response categories of the questions and design layout. Questions are written by the interviewer for clarity as well as for completeness and finally, they are organized on the questionnaire according to the type of survey, respondents and the nature of research questions. While preparing a questionnaire, the researcher also thinks ahead as to how he will record and organize data for analysis. He may conduct pilot testing of the questionnaire with small set of respondents similar to those to be included in final survey.

Step 2: Here the researcher plans how to record data. He also pilot tests the questionnaire with a small set of respondents similar to those in the final survey research. If the researcher has decided to use the interviewers, he trains them with the questionnaire. They ask respondents in the pilot study whether the questions were clear and explore whether their intended meaning was also clear.

Step 3: At this stage, the survey researcher decides on the target population and selects a sample from it. He also decides about the number of respondents to be included in the sample.

Step 4: Here the researcher locates respondents. He uses the desired tools of survey such as mail questionnaire, interview, telephone, etc. He carefully collects data with the help of the selected tool. Every attempt is made to keep the collection of data free from bias that may affect the validity of the research.

Step 5: At this stage, the collected data are analyzed. Before analysis, all data are checked. Nowadays usually computer analysis provides a better and quick means of analysis of the recorded data.

Step 6: This is the final stage where findings are discussed and reported in proper format. Thus, the findings are presented to others for their critical evaluation.

Document or Content Analysis

This method of research has been discussed in Chapter 12.

The Case Study

The Case study is one of the important types of nonexperimental or descriptive research. Case study refers to an in-depth study of one situation or case which may be one subject, group or event (Goode & Hatt 1981; Best & Kahn 1998). Usually the case is a person but a case study is really different from single N experiment in a sense that the former does not involve manipulation of any independent variables. There are researchers who distinguish between a case study, implying a prospective, longitudinal approach and *case history*, implying a retrospective, archival approach.

If we trace the history of the case study method, it becomes obvious that Frederick Le Play (1806–1882) had, for the first time, introduced this method into social science research in his studies of family budgets. Herbert Spencer, an English sociologist (1820–1903) was the first to use case materials in his ethnographic studies. William Healy, a psychiatrist, for the first time adopted

the case study method is his work with juvenile delinquents. Sigmund Freud also used the case study method in the field of psychiatry, especially in his effort to treat his psychoneurotic patients. His famous case history of Sergei Petrov, 'The Wolf Man' published in 1918 under the *title From the History of an Infantile Neurosis* is one of the best classic examples of his use of case study. According to Freud, these case studies confirmed his hypothesis which provided a base to psychoanalysis as a method of treatment. The case study method is also employed by historians, anthropologists, ethnologists and psychologists. Historians have used the case study method for providing descriptive accounts of persons, eras and nations. Anthropologists and ethnologists have used the case study method for making a detailed description and comparison of primitive culture and modern culture. In the field of psychology, exhaustive case studies have been done by Murray and his associates at the Harvard Psychological Clinic. Their case studies included several modes such as interviews, conferences, conversations, dramatic productions, tests of abilities, social reactions to frustration, imaginal productions, etc. There are some evidences which demonstrate that even novelists and wartime correspondents tend to sketch characters of persons and families by using the case study method.

The purpose of the case study method is to understand the important aspects of the life cycle of the unit. In fact, such study deeply analyzes and interprets the interactions between the different factors that influence the change or growth of the unit. Thus, it is basically a longitudinal approach which studies the unit over a period of time. A review of literature in this field reveals that case studies are not confined to the study of individuals and their important behavioural characteristics, rather, case studies have been made of all types of communities and individuals. Whatever the type of individual or community is, the element of typicalness, rather than uniqueness, is the focus of attention in the case study. Therefore, a case of the case study, as Bromley (1986) has suggested, is not only about a 'person' but involves a category of 'individuals'. In the light of this suggestion, the selection of the subjects for the case study needs to be done very carefully so that it may be assured that he or she is the typical of those to whom the generalization is to be made (Best & Kahn 1992).

In case study, data are gathered through several methods or techniques. Some of the important ones are as follows:

(i) Observation of behaviour, characteristics, and social qualities of the unit by the researcher
(ii) Use of questionnaires, opinionnaires, inventories, checklists and other psychological tests
(iii) Analysis of recorded data from newspapers, schools, clinics, courts or other similar sources
(iv) Interviewing the subjects, their friends and relatives, and others

From the aforesaid discussion, the following major features of the case study can be isolated:

(i) The case study is an approach which views a social unit as a whole.

(ii) The social unit need not be an individual only but it may be a family, a social group, a social institution or a community.

(iii) In case study, the unitary character of the social unit is maintained. It means that the social unit, whatever it is, is studied as a whole.

(iv) In case study, the researcher tends to study the aspects of 'What' and 'Why' of the social unit. In other words, here the researcher not only tries to explain the complex behavioural pattern of the social unit but also tries to locate those factors which have given rise to such complex behavioural pattern.

(v) Since case study is a descriptive research, no variables are manipulated here.

(vi) In case study the researcher gathers data usually through methods of observation, interview, questionnaire, opinionnaire and other psychological tests. Analysis of recorded data from newspapers, courts, government agencies and other similar sources is not uncommon.

Based upon the number of individuals, case study may be of two types—the *individual case study*, and the *community case study*. In individual case study, the social unit consists of one individual or person. Since there is only one individual, it emphasizes analysis in depth. Such an individual case study may be fruitful in developing some hypothesis to be tested but it is not useful in making broad generalizations. Such individual case study is a time-honoured procedure in the field of medicine and medical researches. The community case study is one in which the social unit is not a person, rather, a family or a social group. Such case study is a thorough observation and analysis of a group of people who are living together in a particular geographical territory. The community case study tries to deal with different elements of the community life such as location, prevailing economic activity, climate and natural resources, historical development, social structure, life values, health education, religious expression, recreation, impact of outside world, etc. The early community studies have been reported by Lynd and Lynd (1929, 1937), West (1945), Drake and Cayton (1945), Warner and Lunt (1941). Recently, Hollingshead (1949) studied the life of adolescents in the small Illinois Community of Elmtown's Youth. Lucas (1970) has compared the way of social life in the three Canadian communities in *Minetown, Milltown, Railtown: Life in Canadian Communities of Single Industry.*

On the basis of the purpose, a case study may be subdivided into two categories—*deviant case analysis* and *isolated clinical case analysis*. In deviant case analysis, the researcher starts with a difference already found between two persons or groups of persons and his task is to read backward to deduce the condition that might have produced the difference (Warwick & Osherson 1973). In isolated clinical case analysis, the emphasis is on the individual units with respect to some analytical problem. Such a study has been popular in psychoanalysis. Freud's study of little Hans is a well-known case which can be cited as an example of isolated clinical case analysis. Freud's theories of psychoneurosis were formulated through accumulation of many isolated clinical case studies of individuals.

The case study has some *advantages* and *limitations*. Some of the major advantages are as follows:

(i) The case study is a mode of organizing data in terms of some chosen unit such as the persons's life history, the history of a group or society, and some delimited social processes. This has the advantage of intensive study of the social unit.

(ii) According to Goode and Hatt (1981), the case study method provides sufficient basal facts for developing a suitable hypothesis regarding the social unit being studied. This is possible because of the in-depth analysis of the concerned social unit.

(iii) In case study, the researcher gets sufficient fact for making a comparison between two similar social units.

(iv) Goode and Hatt (1981) are of the opinion that the case study provides opportunity for careful examination of all those relevant facts and data on the basis of which a questionnaire or an opinionnaire or any psychological test is to be developed.

Despite these advantages, the case study has some limitations as stated below.

(i) The chief danger in the case study is the response of the researcher itself. Here the researcher comes to feel a false sense of certainty about the conclusions arrived at (Goode & Hatt 1981). The consequences of such feeling of certainty are the temptations to ignore the principles of research design as well as the failure to make explicit just what are the generalizations underlying the analysis of the cases. For example, suppose a researcher has collected 250 cases of drug addiction from different sources. Now he will have a strong temptation to feel that he does have an adequate sample, no matter how much knowledge he has about sampling design. Likewise, he may be in a difficult position to be explicit about the generalization because of the conflicting trends of behaviour of the sample being studied.

(ii) The case study method looks deceptively simple. For effective and proper use of the case study method, it is essential that the researcher must be thoroughly familiar with existing theoretical knowledge of the concerned field and must be skillful in isolating the important variables from many others that are unimportant and irrelevant.

(iii) The case study is a costly method in terms of time and money. It is well known that each case becomes a research in itself and the collection of even 50 cases may consume at least a year. Not only that, there may be a loss of potential cases from the sample drawn due to such a long period of time. Since each case is independently analyzed, it not only consumes time but also a great deal of money of the research fund.

(iv) The subjective bias of the researcher is a constant threat to objective data-gathering in case study. As a consequence, the conclusion loses its dependability and the validity of the study becomes questionable.

(v) The case study is susceptible to what is called post hoc fallacy when the effects are wrongly attributed to factors that are simply associated. In other words, cause and effect relationship is not established in the case study method.

(vi) As longitudinal studies, case studies are confounded by occurrence of many factors and as archival studies, they may provide data with poor reliability and validity.

Despite these limitations, the case study method is a useful method of organizing research observations in social sciences.

Ethnographic Studies

Ethnography is such a nonexperimental or descriptive research which became popular in the latter part of the nineteenth century. It is sometimes known as *cultural anthropology* or more recently as *naturalistic inquiry*. Ethnographic study is a method of field observation or observation of behaviour in natural setting. Originally, it consisted of participant observation, conversation and the use of informants to study the cultural and social characteristics of primitive people. Nowadays, such observation and conversation have been extended to the study of different social groups also. In the beginning ethnographic study was confined to such primitive people like African, South Sea Island and American Indian tribes whose numbers were small and who were geographically and culturally isolated. Major emphasis in such a study was put on language analysis, marriage, child-rearing practices, religious beliefs and practices, social relations, political institutions, etc. In ethnographic study the researchers go to the people of the tribe and data is collected through observation of patterns of action, verbal as well as nonverbal interactions between members of the tribe as well as the researchers and his or her information.

For successful and effective conduct of the ethnographic study, the following three suggestions have been given:

(i) The researcher should personally go to the people of the tribe and live for a long period of time and become an integrated member of the social group.

(ii) The researcher should have the skill to interpret observation in terms of the tribe's concepts, feelings and values while at the same time supplementing his or her own judgement in making an objective interpretation of observation. He should also have learned the native language of the tribe in order to have better adjustment with people.

(iii) The researcher should be trained or at least he should train his informants to record the field data in their own language and cultural perspectives.

The ethnographic study has the following two fundamental assumptions:

(i) The most important behaviour of the individuals in a group or real-life setting is a dynamic process of complex interactions and consists of more than a set of discrete incidents and facts.

(ii) Human behaviour is directly influenced by the setting in which it occurs. The researcher, therefore, must know and understand that setting and the nature of the social structure.

Mead (1928) in her classic ethnographic study, *Coming of Age in Samoa*, studied the development of 53 adolescent girls in Samoan society and compared the same with that of American girls. She reported that there were no basic differences in the physical processes of adolescent growth between Samoan and American girls. Due to cultural restraints, American girls had to face relatively more difficulties during this period.

Recently, ethnographic study has also been conducted over some educational issues thereby shifting its emphasis from the primitive tribe. The study conducted by Morris et al. (1984) is an example. In this study, school principals were observed. In fact, the researchers were interested in determining what principals do and how much time is spent in those activities. Each principal was observed for 12 full workdays. The researchers followed the principal wherever he went and they carefully noted down whom the principal interacted with and by what means, that is, verbally, face-to-face, by telephone, by written words, and the like. They were also careful about the topics which consumed the principal's time and whether spontaneous or planned interactions dominated the conversation. Morris and his colleagues reported in their conclusion that the principals usually spent less than half their workday in offices and that they often acted in a discretionary style in their decision-making. Above all, the principals' behaviour directly affected four major areas: teachers and students, parents and others in the community, superiors and the principals themselves.

Ethnographic study has some advantages and limitations. The important *advantages* are mentioned below.

(i) Ethnographic study is conducted in real-life setting and natural behaviour is observed. The researcher gets inside the minds of the people while at the same time interpreting the behaviour. As such, the facts, got in conclusion, are more dependable and reliable.

(ii) The external validity of ethnographic study is generally high. That is why its generalization is valid and sound.

(iii) The ethnographic study is free from the constraints of more conventional research procedures.

The important *limitations* of ethnographic study are as follows:

(i) For effective conduct of the ethnographic study, a position of neutrality is essential to objective participant observation. Sometimes it is seen that the researchers or their informants fail to maintain the position of neutrality and are overwhelmed by the strong feeling and emotions of the subjects. This defeats the basic purpose of the study and invalidates the objective conclusion of the study.

(ii) The study requires much time and patience on the part of the researchers because they have to live with the people and/or observe their behaviour in a real setting.

(iii) Such a study requires that the researcher must be a trained personnel and capable of interpreting observations in terms of the tribe's concepts, feelings and values while at the same time supplementing his own objective judgement in interpreting observations. In the absence of such qualifications, the researcher is considered not fit for observation.

Despite these limitations, ethnographic study is considered a popular method of research in behavioural sciences.

DIFFERENCE BETWEEN RESEARCH METHOD AND RESEARCH METHODOLOGY

The difference between the two related terms, that is, *Research method* and *Research Methodology* must be clearly understood and explained. By research method, we mean all those methods and techniques which are used in conducting a research. In other words, research methods refer to all those methods which are used by the researchers in conducting the research. Broadly, various research methods used in conducting a research can be put into the following three categories:

(i) In the first category are those research methods that are concerned with the collection of data. These methods are generally used when the data already available is not sufficient enough to arrive at the required solution.

(ii) The second category consists of those statistical techniques which are used for finding a connection between the data and the unknowns.

(iii) The third category consists of all those methods that are commonly used to evaluate the accuracy of the results obtained.

Research Methodology may be understood as the science of studying how the research is conducted. It is mainly a way to systematically solve research problem. For the smooth conduct of research, it is essential that the researcher must know not only the research methods but also the methodology which obviously contains the various steps that are generally adopted by a researcher in studying the research problem with a suitable rationale. The science of research methodology thus specifies that the researcher should not only know how to develop certain tests to be employed, but also know how to calculate various statistics like mean, standard deviation, Pearson *r*, *t*-ratio, *F*-ratio, etc. He should also know which of the various techniques are relevant and which are not, what are the major assumptions underlying the relevant techniques to be used, and the dangers and precautions needed in implementation of those techniques. Research methodology also specifies why a research problem has been undertaken in what way and why the hypothesis has been formulated, why a particular statistical technique for analyzing data has been adapted, etc. Clearly this means that the researcher should design this methodology specific to the problem to be solved because research methodology will differ from problem to problem.

Now a simple comparison between research methods and the research methodology reveals the following features:

(i) The scope of the research methodology is *wider* than that of the research methods. In this way when we talk of the research methodology, we not only talk of the research methods but also consider the rationale behind the methods to be used and also explain why we are using a particular method/technique and why we are not using another method/technique.

(ii) Research methodology contains many dimensions and research methods are simply one of such dimensions.

(iii) Research methodology is broadly a science in itself but research methods consitute simply one aspect of this broad subject.

ETHICAL PROBLEMS IN RESEARCH

The ethical problem in behavioural researches, particularly in the field of psychology, sociology and education, is a vital one because the researchers here conduct experiments mostly on human beings who may be embarrassed, frightened or hurt by the nature of the experiment. To solve such problems some learned agencies like the American Psychological Association, the British Psychological Society, the American Society for the Prevention of Cruelty to Animals (ASPCA) and the United States Office of Education have developed a code of ethics for research with human beings as well as with animals. This code of ethics has been widely welcomed all over the world. Some of the important rules to be observed in animal researches are enumerated below:

1. All animals used in experimentation must be lawfully possessed.

2. Every effort should be made to avoid bodily discomfort to animals subjected to experimentation. Researches, where discomfort to animals is essential, must be done under the proper care of veterinary experts.

3. All animals must be fed properly during the experimentation with due consideration for kind treatment and sanitary environment.

4. Where surgical operations are to be done for fulfilling the purpose of the research, it must be done under suitable general anaesthesia or local anaesthesia. Attempt should also be made to avoid infection throughout experimentation.

5. The animals must be provided with post-operative care in a humane manner.

As we know that research contributes to the advancement of knowledge and ultimately, to the betterment of human beings, it is essential that human beings fully participate and co-operate with the researcher but not at the cost of human rights. Berg (1954) has enunciated three important elements of ethical research with human beings. These elements are *consent*, *confidence* and *standard procedure*. The first ethical consideration is that the person must give his consent to act as a subject. He must give this consent after knowing the risk, pain and discomfort involved in the proposed research. In case the person is mentally ill but is required to act as subject, consent must be obtained from the patient's physician or the patient's guardian. The second ethical consideration, as Berg spells out, is confidence. The human beings who are to act as the subjects, must have confidence in the experimenter in that the latter will not make his attitudes, reactions, and opinions public. It is a natural desire of human beings that others should not know their attitude, opinions and/or performance on a certain task and therefore, it is essential that a human subject be given the confidence that others will not be allowed to know his performances either through the experimenter or through his publication. The third important ethical consideration is that of the standard of acceptable practice. In other words, the procedure that is being followed by the experimenter must be a standard procedure. By standard procedure, we mean that the procedure, must be "tried many times before by many investigators" (Berg 1954). The practice of following standard procedure may sometimes pose a new problem to the experimenter. Suppose the experimenter is trying out a novel and original procedure. In such a situation, the principle of following standard procedure may not be fully satisfied. However, it may be argued in case of a novel and original procedure, which is not standard, that it must be regarded as acceptable by others undertaking the research for the welfare of the subjects. Some of the precautions that a researcher must have regarding human subjects may be enumerated as follows:

1. Where the possibilities for injurious after-effects exist, the researcher should inform the subjects in the beginning and the research should be started only when the subjects agree to participate.

2. When the design of the experiment is such that some intoxicating drugs (which are likely to cause an injurious effect) are to be administered to subjects, adequate facilities must be maintained for appropriate safeguards of the physical and the psychological health of the subjects. Usually, drugs should be administered under the guidance of medical experts.

3. The researcher should safeguard the right to privacy of the subjects. Every subject has a right not to disclose certain information about himself. The use of questionnaire is common in behavioural research. Often, items of the questionnaire relate to religious feelings and personal habits which are not ordinarily disclosed by the subjects. In order to maintain the privacy of the subjects, the researcher should (a) assure the subjects of not making the given information public; and (b) try his best not to put unnecessary and embarrassing questions.

4. The researcher should also honour the right of the subjects to remain anonymous. By this right is meant that personal identities of the subjects should not form the subject of research. For not invading this right of the subjects, the researcher may (a) identify the subjects not by their name but by the numbers assigned, or he may (b) pool the data obtained by a group and report the averages. If the averages are reported naturally, the individual data is ignored and then, the right to remain anonymous is not invaded.

5. The researcher should treat the data obtained from the subjects confidentially. Every subject has a right to insist upon the fact that the data should be treated in a confidential manner and should not be exposed to anyone who wants to see it. Usually, to ensure this right the researcher (a) carries the treatment of the data by numbers rather than by names, and (b) destroys all test protocols when the study is over. Exception exists in the circumstances in which withholding information causes danger to the person or society as well as cases that have subpoenaed records.

6. The researcher should use *deception* wherever it is possible and essential. By deception is meant temporary withholding of information about the hypothesis of the research and other details from the persons who participate in it. In case of some psychological researches it is simply not possible to tell subjects everything which they could be told about the study because if they had knowledge about the actual purpose of the research, they might alter their critical aspects of behaviour which are being studied by the researchers. Alternatively, it is sometimes simply impossible to study a particular psychological process without deliberately misleading the subjects. While the reasoning behind deception is sound, the use of deception also raises some important ethical issues. Is it justified and appropriate for psychologists to withhold information from research participants or even to mislead them? Although this issue remains somewhat controversial, many psychologists believe that deception is *permissible* provided two basic principles are followed.

The first principle involves obtaining informed consent, that is, providing the subjects who are going to participate in the research, as much information as possible about the major events and procedures a study will involve before they agree to participate in it. Such information must be coupled with a statement that they are completely free to leave at any time during the study.

The second principle is called *debriefing* which means that the subjects should be given full information about all aspects of the research study, including deception, after they have participated in it. The purpose of debriefing is to help the subjects leave the study possessing a clear understanding of its purposes and feeling at least as good as they were at the start of the research (Baron 2002). Latest evidences suggest that these two principles, namely, informed consent and debriefing, go a long way towards eliminating the adverse effect of deception (Mann 1994; Sharpe, Adair & Roese 1992). Still, exposure to deception can leave participants with increased feelings of suspicion about what researchers tell them (Epley & Huff 1998).

Obeying these important ethical standards while conducting researches is considered mandatory for psychologists. Violation of these ethical standards may attract punishment by a court of law.

COMPARISON BETWEEN EXPERIMENTAL AND NONEXPERIMENTAL RESEARCH

Social scientific research takes basically two forms—experimental research and nonexperimental research. An experimental research is one where the researcher manipulates some variables and examines their impact under possible controlled conditions. In nonexperimental research the investigator does not have direct control over independent variables, because either their manipulations are already there or they are inherently not manipulable. Whether it is experimental research or nonexperimental research, the researcher tries to verify the postulated relationship between independent variables and dependent ones. Despite this similarity, the following are the major points of distinction between the two:

1. In experimental research, the effect of manipulations in independent variables upon dependent variables is examined. In other words, here the experimenter introduces some changes or manipulations and subsequently, examines its effect. But in nonexperimental research, the situation is reversed. Here the effect is before the investigator and on the basis of that effect, he infers about the possible causes. In other words, in nonexperimental research the investigator, on the basis of dependent variables, infers about the possible independent variables.

2. In experimental research, the researcher has direct control over the independent variables. Whatever the type of manipulation he wants to introduce, he can do so readily and easily. In nonexperimental research the investigator has no direct control over the independent variables. As a consequence, he can't introduce changes of a desired sort in independent variables. The major reason of noncontrol over independent variables is that in nonexperimental research, the independent variables have already been expressed.

3. In experimental research sufficient degree of internal validity is found, whereas in nonexperimental research the level of internal validity is amazingly poor.

4. In experimental research, the investigator can easily assign the subjects randomly into different conditions. Not only this, the various groups of subjects can also be randomly assigned to different treatment conditions. But in nonexperimental research, random assignment of subjects into different treatment conditions does not carry much weight.

5. It is said that in experimental research when an event is brought into the laboratory or in some controlled situation, the nature of the event is changed and becomes different from those which occur in an everyday realistic setting. But in nonexperimental research, events don't change because they are studied in everyday realistic settings.

Thus, experimental research is different from nonexperimental research in many respects.

TYPES OF EXPERIMENT

In social science researches various types of experiments are undertaken, keeping in view the purpose of the experiment. These experiments can be mainly of four types as noted below.

1. *Exploratory experiment*: Exploratory experiment is one which is conducted in a situation in which little knowledge to formulate a possible solution to the problem exists. As a consequence, poorly framed hypotheses (tentative solution) are the only guides for the study. In such experiments, there is no scientific basis for predicting the effect of the independent variable on the dependent variables. Thus, exploratory experiments are broadly conducted at the preliminary stage of investigation and its findings can be the basis for formulating a specific precise hypothesis for further study. For example, if the experimenter observes that most of the hungry subjects learn the given list of nonsense syllables sooner, he may come to the conclusion that hunger facilitates learning. Such findings may be the basis for formulating research hypothesis in future. It is thus obvious that exploratory experiments are conducted mainly to discover whether or not the independent variables are influential in affecting the given dependent variable. According to McGuigan (1990), a common descriptive term for exploratory experiment is "I-wonder-what-would-happen-if-I-did-this."

2. *Confirmatory experiment*: Confirmatory experiment is one in which an explicit hypothesis is subjected to various types of tests and is generally confirmed. According to McGuigan (1990), the common description for a confirmatory experiments is: "I'll-bet-this-would-happen-if-I-did-this." In such experiments, the experimenter clearly formulates a hypothesis and the findings of the experiment are used to test this hypothesis. If the findings support or confirm the hypothesis, the experimenter accepts the hypothesis as true. If the hypothesis is not in accordance with the findings, it is modified to fit the data better and then tested by a new experiment.

3. *Crucial experiment*: A crucial experiment is one that intends to test all possible hypotheses simultaneously. Ideally, a crucial experiment is an experiment whose findings support one hypothesis and rejects all possible alternative hypotheses. Since in any research problem, ordinarily, it is not possible for the experimenter to state all possible alternative hypotheses, therefore, the investigator can't be sure to have a true crucial experiment. However, the concept of crucial experiment is important as an ideal and the accepted view is that efforts should be directed to achieve that ideal.

4. *Pilot experiment*: Pilot experiment, also known as pilot study, is a preliminary experiment, that is, an experiment conducted prior to the major experiment. Broadly, pilot experiment is a sort of dress rehearsal of the major experiments to be undertaken later. Usually, a pilot experiment is conducted with a small number of subjects. It tries to find out the fitness of the procedure being adopted, suggest the values to be assigned to the variables under study, and locate the likely mistakes that might be made in conducting the actual experiment later on. Students should clearly bear in mind that despite the word 'pilot' being generally associated with 'aircraft', 'pilot experiment' has nothing to do with the behaviour of aircraft operators as some students may think so.

A pilot study may be simple or complex. Its degree of complexity rests primarily on the standards, interests and budget of the researcher. It should also be noted that a pilot study may make or break a future research plan of larger scope. In view of what is revealed in the pilot study, a researcher may decide that little or nothing is to be gained through the implementation of a larger study. Of course, sometimes pilot studies show encouraging results and more concrete and sophisticated research plans are implemented subsequently.

There are some specific functions of a pilot study. Some of the important ones are as under:

(i) Pilot studies help in developing better approaches to target populations.
(ii) Pilot studies help the researcher to determine whether or not a more substantial investigation of the same phenomenon is warranted.
(iii) Pilot studies help to discover and ameliorate some problems associated with interviews, questionnaires, and the like.
(iv) Pilot studies help researchers to develop meaningful methods of categorizing the data to be collected.

These are the basic types of experiments usually undertaken in social scientific research. Of these experiments, confirmatory experiments are more important than the other types of experiments.

TYPES OF APPLIED RESEARCH

As we know, applied research is a research designed to offer practical solutions to a concrete problem or address the immediate and specific needs of the society or organizations. There are *three* major types of applied research: *Evaluation research*, *Action research* and *Social impact assessment research*. A discussion on these is given below.

1. Evaluation Research

Evaluation research is the most popular type of applied research. Rutman (1977) used the term evaluation research to describe the evaluation procedures, which used rigorous research methodology. In this type of research the researcher tries to determine how well a programme or policy is working satisfactorily or reaching its goals and objectives correctly. Such research tends to measure the effectiveness of way of doing something. In other words, it aims to assess the effectiveness of different actions in meeting needs or solving problems (Punch 2004). Evaluation researchers use several research techniques (such as survey and field) for evaluating the programme.

There are two common types of evaluation research: *formative evaluation research* and *summative evaluation research*. In formative evaluation research, there is a built-in monitoring or continuous feedback regarding the different aspects of programme being evaluated. A summative evaluation research is one where the researcher looks for final programme outcomes.

Evaluation research is the part of administration of many organizations such as schools, colleges, government agencies, businesses, etc. This type of research greatly expanded in the 1960s in the United States of America when many new federal social programmes were created. In other countries including India, such research is becoming popular day-by-day.

Sometimes ethical and political conflicts often arise in evaluation research because people tend to exhibit opposing interest in the findings about a programme. Results of evaluation research may affect getting a job, building the popularity of a political party, etc. It has been observed that people who are personally displeased with the results of evaluation research often attack the researcher's methods by claiming them to be biased, sloppy, inadequate, etc. Besides these, sometimes evaluation researchers are pressurized to rig a study before they start work.

Evaluation research has some limitations, besides the said difficulties mentioned above. *First,* raw data are rarely publicly available. *Second,* the organization can selectively use or ignore reports of the evaluation research. *Third,* in evaluation reports the focus is narrowed to select inputs and outputs more than the full process by which the programme can affect people's lives.

2. Action Research

The term action research was first coined by Kurt Lewin in 1944. The fields of social psychology, industrial/organizational psychology, education have shown interest in what has been called as action research. Action research is one where focus is given on immediate application rather than on development of theory or on generalization of applications. Here researchers give emphasis upon those problems, which originate in local setting. In fact, its findings are evaluated in terms of local applicability and not in terms of universal validity. Reality is that action research follows two traditions: The British tradition tends to view action research as simply a means of improvement and advancement of practice (Caar & Kemmis 1986) and in the American tradition it is viewed as a systematic collection of data that in a way, provides for social change (Bogdan & Biklen 2007).

An action researcher treats knowledge as a form of power and tends to abolish the line between research and social action. Such researcher assumes that knowledge develops from experience. He also assumes that people can become aware of conditions and learn to take actions for bringing about improvement.

There can be different types of action research and most of them share common characteristics. Some of the common characteristics are as under:

(i) Action research incorporates ordinary or popular knowledge.

(ii) The action researcher tends to raise consciousness or increases awareness by expanding public awareness.

(iii) Those who are studied, participate in the research process.

(iv) Action research tends to focus on power with a goal of empowerment. Such researchers try to equalize power relations between research participants and themselves and avoid having more control, status and authority than the participant.

(v) Action research is tied directly to the actions.

(vi) Since in action research the goal is to improve the conditions and lives of the participants, publishing it in terms of articles and books becomes a secondary matter.

In a nutshell, in action research the participants tend to assume an active role in formulating, disigning and carrying out research as well as the researcher also try to cogenerate knowledge in collaborative processes that continuously incorporate the diverse experience of local groups.

Keeping the above facts in view, Kemmis and McTaggart (2003) have defined action research as collective self-reflective inquiry undertaken by the participants in various social situations for improving productivity, rationality and justice of their own practices as well as their understanding of these practices satisfactorily. Thus action research has a *philosophical element* (concerned with the role of doing or acting in knowing), a *political element* (where focus is on the political aspects of knowledge production) and a *critical social science element* (where the focus is on the empowerment of the participants who are the subject of the research).

3. Social Impact Assessment Research

A type of action research that is more popular in social psychology and sociology, documents the likely consequences for various areas of social life if any major change is introduced in the community (Neuman 2006). Such research aims at evaluating the likely consequences of a planned change. Researchers conducting social impact assessment research examine many outcomes and often work in interdisciplinary spirit. Seven principal areas are assessed by social impact assessment research.

- Community service (such as school enrolments, police responses towards public complaints)
- Economic impact (such as changes in levels of income, failure of profit rate, etc.)
- Social conditions (such as crime rates, elderly people's ability for caring for themselves, etc.)
- Environment (such as changes in noise levels, pollution rate, etc.)
- Demographic consequences (such as population movement, changes in the mix of old and young people in the area).
- Health outcomes (such as changes in the occurrences of diseases or presence of some harmful substances, etc.)
- Psychological well-being (such as changes in the level of stress, self-esteem, self-efficacy, etc.)

Several researches done by different social scientists have revealed that although social impact assessment research is a powerful type of applied research, most of the time its predictions made in different areas have failed to test the truth.

☑ Review Questions

1. Describe the characteristics of psychological research and analyze the various phases of research process.
2. Give a suitable definition of scientific research and explain the nature of scientific enquiry.
3. Compare and contrast laboratory experiment and field experiment. Illustrate your answer with the help of suitable examples.
4. What is 'research'? Discuss the special characteristics of psychological research.
5. Compare and contrast experimental research with nonexperimental research. Give relevant examples.
6. What is 'survey research'? Discuss the relative advantages and disadvantages of the different tools used in survey research.
7. Citing relevant examples, discuss the major steps or stages in psychological research.

8. Citing some examples, discuss the importance of different ethical problems in social scientific research.
9. What do you mean by case study? Discuss its advantages and limitations.
10. What is meant by ethnographic study? Describe its advantages and disadvantages.
11. Distinguish between basic research and applied research.
12. Describe, by means of an example, the major threats to internal validity as emphasized by Campbell & Stanley (1966).
13. Discuss the major reasons for carrying out experiments in natural settings.
14. In what sense can it be claimed that basic research and applied reasearch have a reciprocal relationship to one another?
15. When is a pilot experiment preferred most?
16. Discuss the major types of applied research.

◆

16

SINGLE-SUBJECT EXPERIMENTAL RESEARCH AND SMALL N RESEARCH

CHAPTER PREVIEW

- Meaning and Origin of Single-Subject Experimental Research
- General Procedures of Single-Subject Experimental Research
    - Repeated Measurement
    - Baselines or Operant Level
    - Manipulating Variables
    - Length of Phases
- Basic Designs of Single-Subject Experimental Research
    - Withdrawal Designs
    - Reversal Designs
    - Multiple–Base line Designs
- Data-Collection Strategies in Single–Subject Experimental Research
- Evaluating Data in Single-Subject Experimental Research
- Strengths and Weaknesses of Single-Subject Experimental Research
- Comparison between Single-Subject Research and Large *N* Research
- Small *N* Design: Nature and Historical Perspectives

MEANING AND ORIGIN OF SINGLE-SUBJECT EXPERIMENTAL RESEARCH

The research strategies just described in Chapter 15 were such in which a group or different groups of subjects generally participate in the experiment. In the present chapter, we are going to discuss research strategies in which a single subject ($N = 1$) participates.

Single-subject research, sometimes also referred to as *single case* or *N* of one research, is that in which usually one individual subject participates and the effect of interventions is vigorously studied over the same one individual over time. In other words, a single-subject research is a repeated-measure experiment conducted on one subject. While the focus of single-subject experimental research is the individual subject, no doubt, most of these studies include more than one subject, such as two (Best & Kahn 1998). In this situation it is referred to as *Small N research* (Robinson 1976). Like any other experimental research, in general, single-subject research is a method of testing hypotheses. In general, such research is used to test the hypothesis that a particular treatment will tend to have an effect on one or more behaviours. Single-subject research designs are very useful in clinical research, particularly in the area of behaviour modification and drug evaluation. Most, if not all, behaviour modification research, utilizes the single-subject research design. But it should not be thought that it is limited to behaviour modification research only. In fact, single-subject research design can be applied to a variety of research topics. If we pay attention to single-subject research designs, it will be obvious that these designs are similar to the three quasi-experimental designs, namely, the time-series design, the equivalent samples design and the equivalent materials pre-test-post-test design. The only difference is that these quasi-experimental designs are used with a group of subjects while single-subject research is concerned with individuals.

If we trace the origin of single-subject research design, we come to the conclusion that it developed from the case study method. According to Kazdin (1982), the development of single-subject research design, as it is being currently practised, is largely the outgrowth of the work of B F Skinner who had rigorously conducted animal laboratory research to elaborate what is called operant conditioning. In fact, the research methodology of Skinner, popularly called *experimental analysis of behaviour*, contains many features that are characteristic of the single-subject research design of today. In many of his animal laboratory researches, he has included one or a few subjects upon whom repeated observations of measurements were taken to note subtle changes in behaviour. In the 1950s, several investigators adopted Skinner's methodology to human beings and they successfully indicated the clinical utility of the Skinner experimental analysis of behaviour for autistic and mentally retarded children. As a consequence, a new field called applied behaviour analysis was born with its own journal named *Journal of Applied Behaviour Analysis*, first published in 1968. In this journal, most of the research articles reported the use of single-subject research design.

GENERAL PROCEDURES OF SINGLE-SUBJECT EXPERIMENTAL RESEARCH

There are several aspects of general procedures of single-subject experimental research, which are worth mentioning. Some of these aspects are mentioned below.

1. Repeated measurement
2. Baselines or operant level
3. Manipulating variables
4. Length of phases

A discussion of these procedures follows.

1. Repeated Measurement

One important aspect of single-subject research is the repeated measurement or observation of the same individual over time. The purpose of such repeated measurements or observations is to determine if changes introduced in the experimental condition called *intervention or treatment* affect changes in the subject. The more careful and systematic are the repeated observations, the more valid and reliable data are likely to be gathered.

To assure reliable and valid data, the measurements to be used must be clearly defined. The researcher must be careful in selecting the behaviours to be observed. The behaviours to be observed or measured must be such as the subject, without any hesitation, be ready to exhibit it with reasonable degree of frequency. Sometimes, measurement procedure includes tests, surveys, questionnaires, opinionnaires, attitude scales, etc. The researcher must select such instruments as can be used repeatedly without any contamination or test-interaction effect. For enhancing the reliability and validity of the single-subject research, the measurements must also be done under completely standardized conditions. Standardized conditions in which measurements are to be repeated, include uniformity in the time of day, the circumstances and general surroundings such as location, presence of others, etc. A careful researcher must use the same measurement of observation procedures for each replication of the measurement.

2. Baselines or Operant Level

The baseline, also called operant level, is one of the important aspects of the general procedure of the single-subject experimental research. In this research, a baseline is used to determine the status of the subject's behaviour prior to the intervention by the researchers. Baseline data is often gathered by observing the different aspects of the individual's behaviour. It is to be under study several times prior to the intervention by the researchers. The baseline should be long enough in determining the trend of data. Generally, three types of trends are revealed by the baseline—*a stable rate, an increasing rate* and *a decreasing rate*. For better evaluation of the effectiveness of intervention, the baseline must demonstrate a stable rate. If the baseline is showing an increasing rate or trend, a serious problem in evaluating the effectiveness of the intervention occurs because the baseline, even prior to intervention, is showing a trend in the desired direction. A good baseline is formed only after a minimum of three separate observations; however, five to eight or even more observations will be still better.

3. Manipulating Variables

Like any other experimental research, a single-subject research also requires manipulation of variables. In single-subject research, the ideal condition is that in which one variable should be manipulated at any given time. If two variables are manipulated at the same time, the effect of each can't be separated and the effect becomes uninterpretable. In such a situation, the variables should be manipulated one by one and not all at a time. For example, if the researcher wants to know the effects of systematic desensitization and medication upon an aggressive child, and wants to proceed with a single-subject research design, he should first take any of the two independent variables, that is, systematic desensitization and medication. The researcher may, then, follow the baseline with one intervention or treatment, say, systematic desensitization. After a period of time with systematic desensitization, the treatment or intervention should be removed and the baseline repeated. Following the second baseline, the researcher would introduce the second intervention, that is, medication for some time. After this, the intervention would be removed and the third and final baseline would be introduced. If the researcher manipulated the two variables in the same phase, the effect of each could not have been separated and the effect would remain an uninterpretable one.

4. Length of Phases

In single-subject experimental research, ordinarily, there are three phases—baseline, intervention, baseline. Now the question is: What should be the length of each of these phases? This has been a topic of heated discussion (Bijou et al. 1969; Barlow & Hersen 1984). Barlow & Hersen (1984) have opined that as far as possible, the relative length of each of the different phases should be equal. But this should not be accepted as a general direction for all situations. In fact, there may be a situation in which the first intervention has to be longer than the initial baseline in order to demonstrate an obvious change in behaviour. If this happens, the subsequent phases, that is, the second baseline and the second intervention should be of the same length. However, there is one potential danger in having a longer intervention phase, that is, a carry-over effect is likely to occur. When the intervention phase is made longer, the effect of this intervention continues into the next phase, that is in withdrawal of the baseline. This effect is called *carry-over effect*. Bijou et al. (1969) have recommended short intervention period tends to prevent carry-over effects.

In conducting an experiment based upon a single-subject research design, the above aspects of the general procedure must thoroughly be kept in view.

BASIC DESIGNS OF SINGLE-SUBJECT EXPERIMENTAL RESEARCH

In a single-subject experimental research, the following three types of basic designs are commonly used:

1. Withdrawal Designs
2. Reversal Designs
3. Multiple-Baseline Designs

A discussion of these designs follows.

1. Withdrawal Designs

Withdrawal designs are those in which intervention or experimental treatment, introduced following the baseline period, is withdrawn. There are three basic withdrawal designs: *A-B-A design*, *A-B-A-B design* and *alternating-treatment design*. Before we discuss these three designs, it would be proper to start with the A-B design which is the simplest research strategy in an N = 1 research.

(i) *A-B design*: A-B design is the simplest research design for N = 1 research. In this design a baseline (A) for a behaviour is established and subsequently, it is predicted that the behaviour would continue to exist if no treatment is administered. If following the intervention (B), the behaviour departs from the prediction, the researcher may attribute such behavioural changes to the effect of treatment. The logic of A-B design is that the researcher predicts that behaviour would tend to continue along the same course as the baseline if a treatment or intervention is not introduced. If at the point of intervention the behaviour changes or the response measure noticeably increases, the researcher concludes that his intervention is effective in producing the change. But this conclusion is weak because of the following two reasons:

(a) The researcher does not know what the response rate might have been had no treatment been administered.

(b) The researcher does not know for certainty whether any response change was produced because of the specific intervention or treatment. It might have changed just because the researcher did something different. This is called the *placebo effect* which is one of the important limitations of such designs.

(ii) *A-B-A design*: A-B-A design is a very important and popular research design used in single-subject research. This design has a baseline (A), intervention (B) and baseline (A) sequence. Thus, this design has three phases, each of which represents a series of measurements. In this design the behaviour is studied to examine whether it changes from A (baseline, or control condition) to B, the treatment condition, and whether or not it comes back to the baseline (A) if the treatment or intervention is withdrawn. If the behaviour actually increases during treatment and then again decreases following withdrawal of intervention and thus comes to the level of the baseline (A), a sufficient reason is established for the fact that the response change is a function of manipulation of the independent variable or intervention period. The A-B-A design is more convincing and powerful and yields more reliable and valid data because of the test as to whether or not the withdrawal of intervention does produce a return of the response measure to the baseline.

(iii) *A-B-A-B design*: When intervention is reintroduced after the withdrawal phase, this results in A-B-A-B design which in fact, is analogous to the equivalent time-samples design, a kind of quasi-experimental design. In this design operant level, or baselines (A), and interventions (B), each is repeated twice. According to this design, the behaviour may change from the first baseline (A) to intervention (B), whereupon it may decrease following withdrawal of B (second A) and, finally the behaviour may again increase with the introduction of the intervention (B). As a consequence of this, the functional relationship between the intervention (B) and the behavioural measures is strengthened. Thus, A-B-A-B design provides a better opportunity for careful examination of intervention effect than the simple A-B-A design. Kazdin (1982) has rightly opined that A-B design tends to examine the effect of treatment or intervention by altering the baseline condition (A phase) when no intervention is done, with the intervention condition (B phase). He further says that subsequently A and B phases are repeated to complete the four phases (*A-B-A-B* design). The impact of intervention becomes clear if performance improves during the first intervention phase, and reverts to the original baseline levels of performance when intervention is withdrawn and again, improves when treatment is reinstated in the second intervention. Thus, in A-B-A-B design the investigator observes whether behaviour changes immediately after giving a

treatment variable (first B), whether behaviour reverses when treatment is withdrawn (second A) and whether behaviour improves again when treatment is reintroduced (second B). Since treatment is removed during the second A stage and improvement in behaviour, if any, is likely to be reversed at this point, this design is also sometimes called *reversal design*.

A-B-A-B design can be explained by referring to an experiment reported by Harris, Wolf & Baer (1964). The experiment was conducted on a four-year-old boy who cried a great deal after experiencing even a minor frustration. It was determined that he cried about eight times during each school morning. The response rate (or crying rate) was studied for the first 10 days to establish the baseline (A). The experimental treatment (reinforcement) was introduced for the next 10 days (B). In this experiment, the reinforcement was the special attention from the teacher that the crying brought. For the next 10 days, the teacher withdrew the reinforcement, that is, he ignored the crying episodes of the boy (A). Rather the teacher reinforced some other constructive activities by giving proper attention to the boy. It was found that during this withdrawal period the number of crying episodes sharply decreased and during the last 5 of these 10 days, only one crying response was made. During the next 10 days again, the teacher reinstated the reinforcement by giving special attention to his crying episodes (B) and it was found that approximately original crying response rate was reinstituted. This experiment clearly shows that some given treatment affects the rate of responding. This experiment was replicated with another four-year-old boy, with more or less the same general results.

One common variation of A-B-A-B design is A-B-A-C-A design which is used to assess the effects of two treatments (B and C) relative to the baseline, that is, A. With this design, the researcher can say whether B and C affect behaviour. However, the demerit is that the researcher can't assess their relative efficacies, that is, he can't tell whether one treatment is more effective than the other.

(iv) *Alternating-Treatments design*: This design is, in fact, a subclass of A-B-A-B design. In this design, A and B are the two different treatments that are alternated randomly with a single subject.

Thus, in this design, the treatment A is withdrawn and replaced not by the baseline but by another treatment, that is, B. The basic purpose of this design is to evaluate the relative effictiveness of two or more than two treatments. Suppose the researcher wants to evaluate the relative effectiveness of two different methods of controlling a bad habit of thumb-sucking in a four-year-old child. He may proceed with an extended series of treatments to be alternated in A-B-A-B-A-B-A-B design in which A and B are the two different methods for controlling thumb-sucking behaviour. Over an extended period of time, the researcher may come to the conclusion that method A is better than method B, or vice versa.

It is obvious that in the above alternating-treatment design, no baselines were taken. But an experiment has been conducted using alternating-treatment design in which baselines have been taken. The experiment has been conducted by Ollendick, Shapiro & Barrett (1981). In this experiment, the effect of different treatments for hair-twirling behaviour was evaluated. In this experiment, the effect of two kinds of treatments on hair-twirling behaviour relative to a no-intervention condition was evaluated. As a consequence, there were three conditions—(a) no intervention, (b) positive practice (first intervention), and (c) physical restraint of hair-twirling (second intervention). The baselines were established before any of these three treatments were administered. It was found that there occurred more decrease in hair-twirling in case of first intervention (positive practice) than in case of second intervention (physical restraint).

In modern times, alternating treatment design is gaining more popularity because it has the advantage that treatment is used without a withdrawal and return to baseline.

2. Reversal Designs

This design is also used in single-subject research. In this design, usually two alternative incompatible behaviours are chosen and the researcher establishes the baseline for each behaviour. Subsequently, one behaviour is subjected to one type of experimental treatment and the other alternative and incompatible behaviour is subjected to another type of experimental treatment. When sufficient data has been collected, the treatment conditions are reversed so that the first behaviour receives the treatment initially given to the second behaviour and the second behaviour receives the treatment initially given to the first behaviour. Thus virtually there is reversal of the treatment for the two behaviours. Following McGuigan (1990), one example can be cited like this: Two incompatible behaviours like talking and crying may be chosen for study with this design. First, the baseline for both the behaviours would be established separately. Subsequently, each time the child talks (first behaviour), reinforcement would be administered. However, when the child cries any time during that period, crying (the second behaviour) would not be reinforced (second type of treatment). It is expected that the child's talking behaviour will increase and crying behaviour will decrease. Once this conclusion during this initial phase is reached, the reversal is effected during the second phase in which the child's talking behaviour will not be reinforced but his crying behaviour will be reinforced. It is likely that the reversal of treatments would affect the response rate by increasing crying behaviour and decreasing talking behaviour.

3. Multiple-Baseline Designs

As its name implies, in this design several baselines are simultaneously established prior to the administration of treatments. These designs are basically replication designs. There are three kinds of multiple-baseline designs as mentioned below.

(i) Multiple-Baseline design across behaviours

(ii) Multiple-Baseline design across subjects

(iii) Multiple-Baseline design across conditions or environments

These are discussed below.

(i) *Multiple-Baseline design across behaviours*: In this design, the effect of independent variables across several different behaviours emitted by the same participant is evaluated. The researcher, here, takes several compatible behaviours, that is, the behaviours that occur simultaneously in the individual. He establishes baselines for each behaviour. Subsequently, a treatment is introduced for one target behaviour. If this behaviour changes following the treatment and other behaviours (control) remain stable at the baseline, the researcher concludes that the treatment is affecting behaviour. After some time, the treatment is applied to the second target behaviour of the compatible behaviours and its effects are recorded. In this way, the remaining target behaviours are subjected to treatment one after the other after some time. If the treatment is effective in changing the response rate following its administration to each of the target behaviour of the compatible behaviours, the researcher will have sufficient confidence in concluding about the effectiveness of the treatment.

(ii) *Multiple-Baseline design across subjects*: In this design, a treatment is applied in sequence to the same class of behaviour in different participants who are in the same environment. Another application of this design is that the same treatment is applied in sequence across different behaviours emitted by a single participant. When treatment is applied to the same behaviour of different persons who are in the same environment, sometimes a gap is followed. For example, treatment is applied to one participant after, say, one hour of establishing the baseline to a second after two hours of establishing the baseline, and so on. Because of this time-staggered intervention after establishing the baseline periods, this design has also been called *time-lagged control design*. One example of this design is an experiment reported by Singh & Millichamp (1987). This experiment has been conducted on eight adult females who were residing in an institution for mentally retarded persons. These mentally retarded adults were taught to enhance their independent play by verbal and physical prompts made by the experimenter. First, baselines were established for each female during which they were instructed to play with toys. In fact, two kinds of behaviour, i.e., independent play and inappropriate play were recorded. It was found that for each subject, the baselines for both independent play and inappropriate play were low. Subsequently, a treatment was given to each of the eight participants. During treatment or intervention, both verbal and physical prompts were given, that is, they were verbally instructed to play with the toys kept in front of them and if needed, were provided physical guidance by placing their hands on the toys and comfortably guiding them in their play action. It was found that after such verbal and physical prompts for some time, the amount of independent play increased in all participants and maintained in the follow-up session. The amount of inappropriate play stabilized and remained at low level.

(iii) *Multiple-Baseline design across conditions or environments*: In this design, the treatment is applied to the same behaviour when the participants are in different environments or conditions. To illustrate, the researcher may have four different patients in four different rooms. Different baseline periods may be established for each of these four patients. For example, 20 minutes for one patient, 30 minutes for the second patient, 40 minutes for the third patient and 50 minutes for the fourth patient. After conclusion of each baseline period, the treatment or intervention such as reinforcement may be introduced after making certain similar responses by each patient in their respective rooms. If there is response increment in all patients following the intervention, the treatment is likely to be effective.

The major advantage of the multiple-baseline design is that the researcher has no need to withdraw the treatment once it has been applied. This is in constrast to the withdrawal design.

DATA-COLLECTION STRATEGIES IN SINGLE-SUBJECT EXPERIMENTAL RESEARCH

Like any other experimental research, in a single-subject experimental research two types of behaviour are commonly studied—overt behaviour and covert behaviour. The major data-collection procedure is the observation of an overt behaviour. In single-subject research, there are a number of ways to measure overt behaviour. Some of them are given below.

1. *Frequency measure*: In single-subject research, frequency measure is the popular method of data collection. In this measurement, a simple counting of the number of occurrences of the behaviour that is observed during a given period of time, is made. For example, in an experiment on a mirror-drawing task, the researcher may simply count the number of times the lines drawn by the subject have crossed the boundaries of the star pattern (such counts constitute error). Likewise, if a teacher wants to know how many times a student talks in class without permission, he may simply count the number of occurrences of such behaviour of the child. This is a very simple method of data collection. Sometimes more than one behaviour is counted in this method of data collection.

2. *Duration*: It is a time-based method of data collection in which the actual amount of time during which the individual completes a behaviour, is determined. For example, suppose a particular researcher aims at improving the finger-dexterity skill of a child. So, he will analyze an instructional programme to teach this child to improve the skill and the researcher would like to have the measure of the duration of the finger-dexterity task.

3. *Method of interval recording*: This is also a time-based measure of data collection. It is also known as *time-sampling method*. In this method of data collection, the total observation period is divided into, say, two intervals—observational interval and nonobservational interval. If the researcher is interested in eliminating a particular undesirable behaviour of a child by introducing some intervention, he may observe the child's behaviour for every 4 minutes, for a

2-minute observation period followed by a 2-minute nonobservation period for recording the observed behaviours. Although this method is a popular method of data collection, it has many serious flaws (Barlow & Hersen 1984).

4. *Real-time observation method*: This is an excellent method of data collection in single-subject experimental research. In this method the target behaviours of the study are recorded in their actual frequency, duration and order. Since such detailed recording needs expensive apparatuses and equipment, this method of data collection is not in much use.

In single-subject experimental research, sometimes the behaviours that are studied are not overt. Therefore, the above measures are not employed in data collection. In such a situation, other measures of data collection are used. Three such measures are—

1. *Psychophysiological measure*: In this method of data collection, measures of skin temperature, skin resistance, pulse ratè, blood pressure, etc., are taken with the help of sophisticated instruments.

2. *Self-reports measure*: For knowing the covert behaviour and experiences of the subject, the researchers sometimes adopt the technique of self-report measure in which the subjects are asked to report about their feelings and experiences in the concerned experimental situation.

3. *Response-specific measure*: In this method, data about some specific interval indices of the target behaviour are collected. For example, if a researcher is studying weight reduction, his data might include the count of calories consumed and the distance covered each day by the subject.

These are the popular methods of data collection in single-subject experimental research. Of these various methods, the methods used in case of overt behaviours are common and more popular.

EVALUATING DATA IN SINGLE-SUBJECT EXPERIMENTAL RESEARCH

The data in single-subject experimental research are commonly evaluated through visual inspection. Statistical analysis is rarely used for analyzing the effects of experimental intervention. Through visual inspection, the experimenter or researcher examines such factors as changes in the magnitude and rate of behaviours being studied (Kazdin 1982). According to Kazdin, for assessing the magnitude of the change, the *average rate* of performance and the *level* at the change point should be examined. The average rate of performance is simply the number of occurrences divided by the number of sessions. A change in the level refers to change or shift in performance from the end of one phase to the start of the next phase.

For assessing changes in the rate of behaviour under study, Kazdin emphasizes trend and latency of the response. The trend of the data refers to showing of systematic changes, that is, increase or decrease over time. The latency of change refers to how quickly the change in response occurs after beginning the treatment or withdrawal of treatment, which ensures the effectiveness of the treatment.

Thus, in evaluating the data obtained by single-subject experimental research, the experimenter tends to examine if the average performance changes between phases, if a change in the rate of the behaviour occurs between phases, if the trend of the data as measured by slope of the data lines is in different direction for different phases and how quickly a change occurs after the treatment or withdrawal of treatment is introduced. Some researchers, having poor knowledge regarding evaluation of data of single-subject research, tend to boost the evaluation of such data by high-level parametric and nonparametric statistical analysis. Such practices must not be reported to, as statistical analyses are rarely used in the evaluation of single-subject research.

STRENGTHS AND WEAKNESSES OF SINGLE-SUBJECT EXPERIMENTAL RESEARCH

There are some advantages or strengths of single-subject experimental research. As such, this research is commonly conducted in psychology. Some of the main advantages are as follows:

(i) The biggest advantage of single-subject experimental research is its ability to carry out a scientific investigation with only one subject (or sometimes two). This advantage is more important for a psychologist than for other social scientists, because it saves much time in dealing with many subjects or a large *N* research or between group researches and he is able to have full concentration on only one subject.

(ii) Single-subject research allows the researcher to control the experimental situation more effectively by establishing a good, obvious and continuous measure of the dependent variable (DV) throughout the experimental situation.

(iii) Such researches rarely require statistical tests to be performed for evaluation of data. Only through visual inspection, the experimenter examines such factors as changes in the magnitude and rate of behaviours being studied (Kazdin 1982). This point is a big advantage for those who dislike carrying out statistical computations.

(iv) Both exploratory and descriptive researches can easily be carried out with single-subject design.

(v) Single-subject research allows the researcher to eliminate and hold constant extraneous variables that don't show up until after the investigation is under way. In fact, intra-subject comparison in case of single-subject research provides better control of extraneous variables than the inter-subject comparison, which is the case with large *N* experimental researches. In other words, such research has sufficient degree of internal validity.

Despite these advantages, single-subject research has some *disadvantages* or *limitations*. Some of these disadvantages are as follows:

(i) *Inappropriateness*: Single-subject research is not appropriate for certain types of psychological research such as survey or ex post facto situation.

(ii) *Practical limitations*:. Single-subject research is time-consuming research. It generally takes several fortnights or months to complete whereas many large *N* research designs can be carried out in only one session. In fact, single-subject research is feasible only when the subject remains willing to co-operate by giving sufficient time for fortnights or months. Sometimes it has been reported that the treatment or intervention in such research is difficult to be measured repeatedly or the nature of DV is such that it can't be measured repeatedly. For example, once a problem is solved, there is no further DV measure to be taken.

(iii) *Order effects*: In single-subject research, there are obvious order effects which result in confounding and limiting the quality of generalization. Since there is only one subject over whom treatments are applied, it may be that practice of one treatment may be carried over to the later treatment and may improve performance under later treatment or the first treatment may introduce fatigue to the second treatment. Still another type of order effect may be in limiting the generalization. Suppose, for example, the investigator is evaluating the effectiveness of a particular therapy. Such therapy could be ineffective in the first B phase of A-B-A-B design but could be effective in second B phase. From this, it may be erroneously concluded that for successful application of therapy, such therapy should be presented, then withdrawn and then, again presented.

(iv) *Baseline problems*: Barlow & Hersen (1984) have discussed many problems related to baselines in single-subject research. One of the basic assumptions of single-subject research is that a stable baseline has been established. If the baseline varies, it is difficult for the researcher to conclude a reliable change in behaviour following treatment or intervention. Not only this, sometimes in A-B-A-B design which has two baselines (A), the second baseline becomes markedly different from the first. In such a situation, the researcher would not be able to assess the effects of intervention with certainty.

(v) *Irreversible effects*: Another disadvantage of the single-subject research is that some treatments given to the subjects are not reversible. For example, in drug evaluation research the impact of some of the drugs may linger up to the return-to-the baseline phase because of its profound effect. In such a situation the effect of intervention is difficult to be isolated. Not only this, sometimes the effect is reversible; however, such reversal is not done because that may be unethical and/or practically difficult. Suppose the researcher is giving therapy to a phobic patient following A-B-A design. The therapy has been given in B phase and he is able to eliminate the irrational fear. But in second A phase the design requires to return to the baseline and in this case, phobic fear is again introduced in patient. Obviously, this can be done by the therapist or researcher. But will such reversal be ethical? Certainly not.

(vi) *Lack of effectiveness of treatment*: In single-subject research, a doubt regarding the effectiveness of treatment or intervention introduced often occurs in mind. Are changes in the behaviour of the subject due to the introduction of IV (or a treatment)? Or, is it merely because something different from the baseline is being done? Nothing definite can be postulated. One solution often suggested to this limitation is to use A-B-C-B design and consequently if behaviour during the C phase very much resembles the behaviour during the A phase, the probability is that the change during the B phase is specifically produced by the intervention introduced.

(vii) *Researcher's bias*: Some biases on the part of a researcher are often seen in introducing the intervention or treatment. Generally, the researcher waits for some cues indicating change in subjects' behaviour and then he introduces a withdrawal of the treatment. Such practices by the researcher are not scientific. To remove this limitation, the researcher should precisely specify the length of the baseline and treatment phase before data collection, and those specified periods should be rigidly followed.

(viii) *Dependent behaviours*: In single-subject research design, particularly in multiple-baseline design across behaviour, the researcher assumes that different responses or behaviours being studied are independent. Consequently, the treatment that affects one behaviour but not the other is considered to be an effective one. But if the behaviours are correlated, the treatment that affects one baseline tends to affect the other, even though the treatment is not applied to those other responses. In such a situation, the interpretation becomes difficult. One interpretation may be the generalized effect of treatment indicating response generalization, that is, the intervention first affected one response and then the other through the process of generalization. Yet another interpretation may be that the treatment intervention had generalized effects on all behaviours.

(ix) *Lack of external validity*: Single-subject research in comparison to the 'between groups research' or large *N* research has poor degree of external validity or generalizability. Since such an experiment is conducted only on one subject, the conclusions obtained are difficult to be generalized to other participants. But if single-subject research is replicated systematically with several other participants or the different researchers replicate the study, the generalization can be made with sufficient confidence.

(x) *Magnitude of effects*: In single-subject research design where repeated measurements are taken to the same subject, it is very difficult to determine whether the treatment is effective for naive individuals who have not experienced those previous treatments. When subjects are exposed to previous conditions, the researcher can easily estimate whether a treatment is effective. But since a change in the behaviour following that treatment has followed some previous treatment, the researcher can't tell how effective the treatment is for individuals who have not experienced the previous treatment.

Thus, we find that single-subject research has some limitations, despite some of its advantages. Still single-subject research has been a popular method of research in the *Experimental Analysis of Behaviour* as well as in other areas of research such as drug evaluation and behaviour therapy.

COMPARISON BETWEEN SINGLE-SUBJECT RESEARCH AND LARGE *N* RESEARCH

Single-subject research and large *N* research are both experimental researches where IVs are manipulated and its effect upon DV is examined. However, there are major points on which the two differ.

1. *Control techniques*: Large *N* research generally employs randomization, making an extraneous variable an independent variable, and various statistical controls whereas single-subject research does not. Single-subject research does use randomization as a means of subject-selection but not as a technique of controlling extraneous variables in the experiment. Although both large *N* and single *N* or small *N* experiments employ elimination and constancy as control techniques, the ways they use them differ. In large *N* experiments, extraneous variables are eliminated or held constant at the beginning of the investigation. In single *N* experiments elimination and constancy are employed as the investigation is carried out, as also at the outset. In such experiments, the experimenter not only tries to hold the extraneous variables constant but also increases his control over the experimental situation by setting continuous measures of dependent variables. We also find that large *N* experiments control chance fluctuations by using a large number of subjects whereas single *N* experiments control them by using an increased number of responses of the subject.

2. *Manipulation of dependent variables*: In large *N* experiments the experimenter generally sets up his experimental design, controls the extraneous variables as far as possible, applies an independent variable and takes a one-shot measure of the dependent variable (or DV). Here little is done about DV except this one-shot measure. But with single *N* research, the experimenter can actually change the DV by varying the extraneous variables. In fact, this procedure is carried out in an effort to develop a stable DV measure prior to applying the independent variable.

3. *Monitoring the experimental data*: In large *N* experiments, the experimenter makes little effort to monitor and analyze the data until the experiment is over. But in single *N* experiment, the experimenter continually monitors his data so that a watch can be kept on changes in extraneous variables that may occur during the experiment itself. Such monitoring also becomes necessary because shifting from one condition to the other is usually determined by the behaviour of the subject.

4. *Data analysis*: In large *N* experiments the experimenter variously uses statistical tests like t test, F test, Pearson r for determining the significance of the data. Frequently in single *N* or small N experiment, no statistical test is used for evaluating the data. Very simple measures like averages are computed or the data are simply presented in a graphic form for overall comparison of the subject's behaviour from one session to another.

5. *External validity of the results*: External validity refers to the generality of the results. In large *N* experiments, the experimenter tends to increase the generality of the results by increasing the number of subjects in the investigation. Many researchers believe that simply by increasing the number of subjects, the generality of results can be increased because the obtained mean and variance of the sample will closely approach the true mean and variance of the population. But this, in fact, is not the correct assumption. Suppose an experimenter finds that a particular method of teaching is effective on a fifth grade student. Can the possibility of making this method effective be generalized to a mentally retarded child simply by adding additional normal children of the fifth grade? The answer will be definitely 'No'. In fact, the procedure of generality from one

type of situation to another can be carried out through inductive reasoning. In other words, by comparing the similarities and differences between the two situations, the experimenter determines how generalizable the data is. Single-subject experiment tends to determine the generality of results by this inductive reasoning rather than by inferential statistical methods as we find in case of large N experiments.

Thus, single-subject research differs considerably from large *N* research in various important aspects such as employing control technique, data analysis, generality of data, etc.

SMALL *N* DESIGN: NATURE AND HISTORICAL PERSPECTIVES

At the time psychology emerged as a new science in the second half of the 19th century, statistical analysis was also in its infancy. Galton was planning to conceptualize correlation techniques and higher inferential statistics were used only after Fisher's work on Analysis of Variance (ANOVA) appeared in the 1930s. Before this, small *N* studies prevailed in psychology. As its name implies, researches based on small *N* designs are those in which more than one subject, that is, a small number of subjects participate.

Some of psychology's pioneers like Hermann Ebbinghaus, Charles Darwin and Watson and Raynor used the smallest *N* possible, that is, they studied their own behaviour or the behaviour of a single individual. Ebbinghaus conducted exhaustive study of his own ability to learn and recall lists of nonsense syllables. Charles Darwin conducted study on child development by keeping a detailed activities of his own son's childhood. Watson and Raynor conducted study on little Albert.

At Leipzig University in Wundt's laboratory, experiments based on small *N* designs were very popular. These studies normally involved a very small number of research participants who were mostly doctorate students. For example, the bulk of J M Cattell's dissertation research on reaction time included data from just two persons—Cattell and his fellow student Gustav Berger. In 1893, Dressler also conducted his famous study of facial vision using only three participants. Although small *N* studies and many similar other studies were very popular, large *N* studies were not completely absent. Such large *N* studies could be found in educational psychology and in child study research (Danziger 1985)

Reasons for Small *N* Design

There are several reasons why small *N* studies are used. Some of the important ones are as under:

(i) Modern advocates of research featuring one or just a few subjects argue that grouping data can obscure individual performance and produce misleading results. The process of summarizing data from large groups of individuals sometimes produce results that, in fact, can characterize no single individual who took part in the study. This viewpoint has been strongly argued by Sidman in his book entitled *Tactics of Scientific Research* (1960). Since group averages are contaminated by individual differences between subjects, Sidman (1960: 274) has rightly opined, "group data may often describe a process or a functional relation that has no validity for any individual."

(ii) There are some practical problems with researches based on large *N* design. Such researches sometimes become very cumbersome because potential subjects are difficult to find. This happens especially in the field of abnormal and clinical psychology where the researcher wants to study a particular disorder. Suppose the researcher wants to test the effectiveness of new therapy for those who are suffering from hypochondriasis. Here the simplest large *N* design would compare two groups of participants with perhaps 20 per group. Matching would be necessary, probably for characteristics like gender, intelligence, socio-economic status, etc. For obtaining two matched groups, a relatively large initial sample (more than 100) of people suffering from hypochondriasis is needed. Now the researcher may not be getting such large sample or he will have to wait for a longer period of time.

A closely related problem also occurs in some animal research, especially if surgery is required. The procedure of surgery is time consuming and expensive and the animal colony itself can be difficult to maintain. In some animals being studied, it might be hard to obtain them or may require long training. For example, a research on learning sign language to chimpanzees and other apes may require hundred of hours per animal. If large *N* design is used, many such animals would be required and this may prove to be a Hereculean task for the researcher. The study conducted by Patterson and Linden (1981) has shown that during teaching sign language to a lowland gorilla, the ape could learn more than 400 different signs at the age of 10. The study had began when the ape was just a year old.

A still another related reason of why small *N* research should be preferred has been provided by Skinner who believed that the best way to understand, predict and control behaviour is to study single individual intensively. The researcher can derive general principles only after exhaustive study of individual cases. In simple words, psychology should be an inductive science, reasoning from specific cases to general laws of behaviour. As Skinner (1966, 21) has opined, the researcher should "study one rat for a thousand hours" rather than, "a thousand rats for an hour each or a hundred rats for ten hours each".

Thus large *N* researches may occasionally fail to reflect the behaviour of the individuals and they may not be feasible even if they are desired. This naturally paves the way for small *N* researches.

☑ Review Questions

1. What is meant by single-subject experimental research? Discuss the general procedures of this research.
2. Discuss the different important designs of single-subject experimental research.
3. Outline the important data-collection strategies in single-subject experimental research. How will you evaluate the data obtained in such researches?
4. Point out the major strengths and weaknesses of single-subject experimental research.
5. Make a comparison between single-subject experimental research and large *N* experimental research.
6. Under what conditions is single-subject experimental research considered more appropriate than a multiple-group experimental research?
7. What types of research questions can't be easily answered by conducting single-subject experimental research?
8. How will you defend the criticism that a single-subject experimental research lacks external validity?
9. Why do we call A-B-A-B design as a reversal design?
10. Distinguish between baseline and intervention stage of single-subject experimental research.
11. Discuss the probable reasons for small *N* design.

◆

17

HISTORICAL RESEARCH

CHAPTER PREVIEW

- Meaning of Historical Research and Its Necessity
- Steps in Historical Research
- Sources of Historical Data
    - Primary Sources
    - Secondary Sources
- Historical Criticism
    - External Criticism
    - Internal Criticism
- Limitations of Historical Research

MEANING OF HISTORICAL RESEARCH AND ITS NECESSITY

Historical research is very different from the research conducted by most of the behavioural scientists. Kerlinger (1986, 620) has defined historical research as "the critical investigation of events, developments and experiences of the past, the careful weighing of evidence of the validity of the sources of information on the past, and the interpretation of the weighted evidences". Thus, historical research refers to the application of scientific method to the description and analysis of past events. In fact, history—the meaningful record of past human achievements, deeds and misdeeds—helps us understand the present and, to some extent, predict the future course of events. It should not be thought that history is merely a list of chronological events. Rather, it is supposed to be a factual integrated account of the relationships between persons, times and events. Historical analysis can be directed towards an individual, an idea, a movement or an institution. But none of these objects of historical observation can be treated in isolation because of the intense interrelationship among institutions, movements and man.

Historical research becomes necessary not only for knowing the past but also for understanding the present and predicting the future in that context. Substantiating the role of past, one historian Arthur Schlesingen rightly commented, "No individual, let alone a social scientist, can ignore the long arm of past." The noted thinker, George Bernard Shaw, also commented in much the same vein, "The past is not behind the group; it is within the group."

There are some social scientists who try to establish link between the past experiences on the one hand and the present attitudes and values on the other. Among these social scientists, the name of W I Thomas, well known for his concept of "the processing of becoming", and A T Adorno et al. (1950), who are the authors of *Authoritarian Personality*, are important. Adorno et al. vigorously demonstrated that the roots of prejudice were in early family experiences. He examined various techniques of combating such prejudices at the source. The Committee on Historiography of the Social Science Research Council also concluded that the analysis of dynamics of change and development was now regarded as a fruitful step towards understanding the relationship of man, institution and movement. Anthropologists also utilized the historical data for studying a culture group. In fact, a historical investigator like other investigators, collects data, evaluates data for validity and interprets the data using historical method or *historiography*.

The relevance of historical data for social research is great in the following three types of situations:

(a) When they are presented as a complex phenomenon of social forces or social dynamics

(b) When social phenomena reflect complex social processes in a meaningful way

(c) When various dimensions of interrelationships such as psychological, economic, educational, religious and political make significant contribution towards a unified complex pattern

Historical research may be qualitative or quantitative or both depending upon the nature of problem and the type of data processed (Best & Khan 2006)

STEPS IN HISTORICAL RESEARCH

Like a behavioural researcher, a historical researcher also follows some steps in sequence. The following are the major steps:

(i) The first step in historical research is the selection of a problem. All problems are not suited for historical research. Only those problems are to be picked up for the study which need scanning of historical records and have some social utility. No current problem can be meaningfully studied with the help of historical research.

(ii) The second step will be formulation of some hypotheses. Hypotheses are explicitly formulated in historical investigations of education. The historian gathers evidences and then carefully evaluates the trustworthiness of the hypothesis. If the evidence is compatible with the hypothesis or its consequences, that hypothesis is confirmed. It is through the synthesis of hypothesis that historical generalizations are ultimately established.

(iii) The third step is collection of data. Historical data are collected with the help of primary source or secondary source. Primary sources are eyewitness accounts reported by actual observers or participants in the event. Secondary sources are the accounts of an event that are gathered after the reporter has talked with the actual observer or has seen the accounts of an observer.

(iv) At this step, the data is analyzed and a generalization is arrived at. Precautions are taken to see that all the data have been collected and that nothing has been left out. In reaching conclusions, historians employ the principles of probability as is done by a physical scientist. They rigorously subject the evidences to critical analysis in order to establish perfect authenticity and correctness.

SOURCES OF HISTORICAL DATA

Like any other research, collection of data is an important aspect of historical research. There are two major sources of data collection in historical research.

(1) Primary Sources

Primary sources are eyewitness accounts of events reported by an actual observer or participant in an event. Such important sources, which are commonly used, are as mentioned below.

(a) Documents

Documents are those records which are kept and written by actual participants or direct witnesses of an event. These sources are also produced for the purpose of transmitting information to be used in the future. Documents which are included as primary sources are official minutes or records, constitutions, certificates, declarations, licenses, deeds, wills, affidavits, books, diagrams, paintings, pictures, newspapers, catalogues, films, advertisements, maps, handbills, recordings, findings of research, reports, and so on.

Direct use of documents proves more useful:

(i) When events, which these documents depict, have not yet been analyzed by the historians

(ii) When such events have not yet been incorporated into the writings of the cultural historical settings

(iii) When the purpose is to verify certain events directly

(iv) When certain aspects of life in which the investigator is interested have not been embodied in later writings of analytical historians

(v) When there is a missing link in knowledge that needs to be connected so that the whole event may be meaningfully understood in a complex whole

(vi) When complete chronology of events is lacking

(vii) When some controversial points need to be settled

One precaution in the use of documents is that for gaining perspective, they should not be preferably used until they are at least a few generations old (Young1992). But contemporary documents may be the only source of information available at the time and the closer a person remains to the time, spirit and scene of action, the more able he will be in understanding the events.

(b) Remains or relics

Remains or relics are objects which are associated with some persons, groups or periods. Fossils, skeletons, tools, utensils, weapons, clothing, buildings, coins, furniture, pictures and art objects are examples of such remains or relics, which, though not directly intended for use in transmitting information, provide a clear evidence of the past life and happenings.

(c) Oral testimony

Oral testimony is the spoken account of a witness or a participant in an event. Such evidences can be obtained in a personal interview or can be recorded or transcribed as the witness or participant relates his or her experiences.

(2) Secondary Sources

Secondary sources are accounts of an event that were not actually witnessed by the participant or the reporter. The writer of the secondary sources is not on the actual scene of the event, rather he merely reports what the person, who was there, said or wrote. Many books on history and encyclopedia are examples of secondary sources because thay are often far from the original contents. Secondary sources of data are used less than the primary sources because of the errors that may result when the information is passed from one person to another.

Some types of material may be used as primary sources for some purposes and secondary sources for some other purposes. For example, a high-school textbook on Indian history is ordinarily a secondary source. But if the investigator is studying the gradually changing emphasis on nationalism in high-school textbooks on Indian history, the book would be a primary source of document or data.

HISTORICAL CRITICISM

Historians are often in difficulty in getting trustworthy and authenticated data because they ordinarily don't use the method of observation. Since past events can't be repeated at the will of the investigators or historians, they are forced to depend upon those who witnessed or participated in these events. They are required to make a careful analysis so as to distinguish unerringly between a true and a false event or between relevant and irrelevant information. The data which are trustworthy, authentic and worth using are called historical evidences. A historical evidence is derived from historical data by the process of criticism which is of two types—external and internal.

(1) External Criticism

External criticism is one which establishes the authenticity or genuineness of the historical data. Historians employ various tests of genuineness of data. For this purpose, they thoroughly examine a signature, handwriting, script, language documentation, knowledge available at that period as also what is known. Sometimes historians resort to physical and chemical tests of ink, paint, paper, cloth, metals or wood. They try to establish whether or not these elements are consistent with known facts about the person, knowledge available and the technology of the period from which the remains have been obtained.

(2) Internal Criticism

Internal criticism is one in which the historian tries to evaluate accuracy or worth of the data. External criticism reveals only the genuineness of the data but whether or not these data reveal a true picture is indicated only by internal criticism. Here historians try to establish whether the writers or discoverers of the data were honest, unbiased and actually acquainted with facts or whether they were themselves biased or too antagonistic or too sympathetic to a true picture, whether they were in agreement with other competent authority of that period, whether they wrote about the events freely or were subject to pressure, fear or vanity. Although these questions are difficult to be answered, historians try to establish that the data are not only authentic but also accurate ones, failing which they can't introduce historical evidence worth consideration.

LIMITATIONS OF HISTORICAL RESEARCH

Historical method of social research has some limitations which are even accepted by historians. Some of these limitations are given below.

(i) One common limitation is whether or not the data being made available are dependable.The problem becomes serious when the data relate to a distant past. Often modern historians agree that a few narrators of historical events report them without injecting therein some such impressions of their own—a fact that introduces an impressionistic and propagandistic element.

(ii) No life-size writing of historical events are possible. Each historian records what he considers to be relevant and ignores what he considers irrelevant. Accordingly, no life-size history can be written. This introduces an element of bias in historical research.

(iii) Not all the happenings in time and space can be known at the time of writing. Historians try their best to collect maximum information from all sources. Despite their honest intention, some happenings, especially those in distant villages or places, may not come to the notice of historians. As such their record may remain incomplete and only partially dependable.

(iv) Some personal biases and private interpretations often enter unconsciously into the interpretation of historical events. Historians on the one hand, often find it imperative to omit a mass of detail, and on the other hand, in the remaining mass they retain, they often interject their own conclusions which are often deviated and biased ones.

(v) In historical research, sometimes difficulties arise in recognizing the problem. Often it is seen that historians state too broad a problem which is not worth investigating. Therefore, historians must select a penetrating analysis of a limited problem rather than a superficial sketch of only some broad area.

Thus, we see that historical researches have some limitations which tend to make the generalizations vague and the conclusions inappropriate ones.

☑ Review Questions

1. Discuss the major sources of data in a historical research. Also, point out the major limitations of a historical research.
2. What is meant by 'historical research'? Discuss the steps involved in a historical research.
3. Point out the relevance of historical criticisms in historical researches.

◆

18

THE PROBLEM AND THE HYPOTHESIS

CHAPTER PREVIEW

- Meaning and Characteristics of a Research Problem
- Sources of Stating a Research Problem
- Important Considerations in Selecting a Research Problem
- Ways in Which a Research Problem is Manifested
- Types of Research Problems
- Importance of Formulating a Research Problem
- Steps in Formulating a Research Problem
- Meaning and Characteristics of a Good Hypothesis
- Formulating a Hypothesis
- Ways of Stating a Hypothesis
- Types of Hypotheses
- Sources of Hypotheses
- Functions of Hypotheses

MEANING AND CHARACTERISTICS OF A RESEARCH PROBLEM

Any scientific investigation starts when a person has collected many facts but all that can be said on the basis of those facts is that there is something which we do not know. Here, a problem originates. A problem is manifested in at least three ways. *First,* when there is a clear gap in the results of several investigations in the same field, a problem is said to exist. *Second,* when the results of several investigations disagree with each other, a problem is manifested. *Third,* when the facts in any field are found in terms of unexplained information, a problem is said to exist. The statement of a research problem is often a difficult task for the researcher. Sometimes he has only a diffused and even confused idea regarding the problem. Before selecting and formulating a problem, he must know what is meant by a *research problem* or a *problem.* A problem statement may be defined as an interrogative testable statement which expresses the relationship between two or more than two variables. Analyzing this simple definition of a problem statement, it can be said that there are three important characteristics of a problem statement:

1. A problem statement is written clearly and unambiguously, usually in question form. A few examples of problems may be:

> What is the relationship between IQ and classroom achievement?
>
> What is the relationship between anxiety and intelligence among school children?
>
> Do students learn more from a democratic teacher than from an authoritarian teacher?

All the above examples state a problem (to be investigated) in question form.

Research questions, whether they are prespecified or developed during the research work/project, are central and serve the following functions:

(i) They organize the project and provide it certain direction and coherence.

(ii) They delimit the project and display its boundaries.
(iii) They keep the researcher focused and attentive during the project.
(iv) They also point about the data which will be needed.
(v) They provide a general framework for writing up the project.

Researchers have revealed some characteristics of *good* and *bad* research questions. A few important characteristics of a good research question are as under:

(i) *A good research question is clear*: A good research question is unambiguous and easily understood.

(ii) *A good research question is specific*: A good research question possesses such concepts which are specific to the extent that they are connected to data indicators.

(iii) *A good research question is answerable*: A good research question clearly shows that what data are required to answer them and how the data will be obtained.

(iv) *A good research question is interconnected*: Good research questions are related to each other in a meaningful way. They are in no way unconnected.

(v) *A good research question is substantively relevant*: A good research question is an interesting and worthwhile question for the research effort.

Bad questions, on the other hand, are questions which fail to satisfy one or more of the above criteria. Mostly this happens because such questions are unclear and not specific.

2. A problem statement expresses the relationship between two or more than two variables. This kind of problem, which expresses the relationship between two or more variables, permits the investigator to manipulate one of these variables and examine its effect upon the other variable. Such a problem is completely different from a problem in a descriptive study where the investigator cannot manipulate a variable, rather he simply observes or counts the occurrence of a particular variable. For example, if the investigator puts the question, "How many students have an IQ above 110 in Class VII?", it will be an example of a problem in which he is required not to manipulate, but simply to count the students having an IQ higher than 110. This illustrates the problem found in a descriptive study. But suppose he puts the problem like this: "Is higher IQ related to classroom achievement?", it illustrates the problem found in a scientific study because the problem statement explores the relationship between IQ and classroom achievement.

3. A problem statement should be testable by empirical methods. In other words, a problem statement should be such that it should be tested through the collection of data. A problem statement in which the stated relationship between variables cannot be tested is not a scientific problem.

Apart from these characteristics, there are three additional characteristics of a problem statement.

First, a problem statement should avoid moral or ethical judgements simply because such statements are very difficult to study. For example, questions like the following should be avoided: Should a widow remarry? Should children avoid cheating in all situations? Should a person not tell a lie in all situations?

Second, the problem should be of sufficient importance. In other words, the problem should not be too trivial, or too expensive in terms of time, money and effort, otherwise the investigator will not be able to furnish any answer immediately or even in the future.

Third, problems can be general or specific. Scientifically, a too general or broad problem is quite useless because it cannot be tested. Therefore, the problem must be reduced to a workable size or it must be specific. But a problem should not to be too specific because "great specificity is perhaps a worse danger than too great generality" (Kerlinger 1973, 24). A too specific problem is, therefore, a trivial problem.

SOURCES OF STATING A RESEARCH PROBLEM

Stating a suitable problem is always a difficult task for the researcher. To solve this difficulty, it is essential that the researcher acknowledges the likely sources to which he may go for formulating a suitable research problem. These sources are as follows:

1. A careful examination of problems which are being faced by teachers, students, guardians may throw light upon several topics that are worth investigating. For example, student unrest is one of the most important problems which school and college teachers are facing today. Thoughtful deliberations may enlighten several problems worth investigation, e.g., what type of students participate in unrest? In which age group is the unrest most obvious? To what extent are teachers and guardians responsible for student unrest?

2. Those who are looking for a suitable problem should consult textbooks and research journals. Many research articles suggest problems for further investigation. Not only this, they also demonstrate the use of different techniques and procedures to be followed in investigating a certain problem.

3. Problem seekers may also consult research professors, instructors and advisors who are regarded as the most competent persons for suggesting a suitable problem.

4. Social changes and educational innovations constantly bring forth new problems which are worth investigating. These new problems pave the way for new opportunities for further researches. The problem seeker may find a suitable research problem grown out of the various social changes and educational innovations.

Best & Kahn (1998) have outlined several different sources of locating a research problem. Among these, the important ones may be outlined as mentioned below.

(1) Programmed instruction
(2) Television instruction
(3) Team teaching
(4) Homework policies and practices
(5) Field trips
(6) Extracurricular programmes
(7) Student's out-of-school activities: employment; recreation; cultural activity; reading and television viewing
(8) Teacher's out-of-school activities: employment; political activity and recreation
(9) Open classroom
(10) Multiple textbooks
(11) Independent study programmes
(12) Educational organizations
(13) In-science programmes
(14) Sex education
(15) Nonschool-sponsored social organizations or clubs
(16) Attribution of success and failure
(17) Special education
(18) Comparison of the effectiveness of two teaching methods/procedures
(19) Factors associated with selection of teaching/nursing/social work as a career
(20) Case studies
(21) Socio-economic studies and academic achievement
(22) Administrative leadership

(23) Stress and achievement

(24) Self-image analysis

(25) Vocational objectives of students

The above topics are the general sources of locating a problem. Therefore, a great deal of refinement must necessarily be done in order to carve out a research problem from them.

Young (1992) has suggested the following three major general sources of locating a research problem in the field of behavioural sciences:

(1) *Documentary sources*: These sources include the analysis and inspection of official and unofficial statistics, local newspaper accounts, census publications from which different types of descriptive materials are easily derived.

(2) *Personal sources*: These sources include examination of some professionals who have right knowledge and insight into the data desired as well as the interview of those persons who are directly involved in the commission of the concerned act or behaviour.

(3) *Library sources*: Such sources include study of various textbooks, journals, monographs and newspaper analysis.

These sources supply both practical and theoretical knowledge and help a lot in accumulating pertinent data regarding formulation of a research problem.

Obviously, then, there are different sources from which a research problem can be finally traced.

IMPORTANT CONSIDERATIONS IN SELECTING A RESEARCH PROBLEM

Before the selected research problem is considered appropriate, and before the actual investigation starts, several questions should be raised. When the questions are positively answered, we can regard the problem as a good scientific problem and investigation can be started.

1. Is the problem significant? Are there enough variables to be investigated? Will the solution of the problem make a significant difference in current psychological, educational and social theories or practices?

2. Can the problem be solved by the process of research? Is the problem such that relevant data can be gathered and an answer to the problem be found?

3. Is the problem a new one? If the problem is such that it has already been investigated, a researcher will fruitlessly spend time and money in replicating the study. Therefore, the problem should be an original and novel one, although there are times when it becomes necessary to replicate a previous study so that the conclusion can be verified or the findings can be generalized to a broader situation.

4. Has the problem theoretical value? Will the problem fill the gap in the literature? Will it contribute to the development of a theory?

5. Is the problem workable? A problem may be a good one but it may not suit the researcher. The researcher should get the answers to the following questions in the affirmative: Am I competent to carry out the research on the selected problem? Will I have enough time to complete the research? Will I have adequate financial resources to complete the research? Will I be getting the cooperation of subjects on the proposed research problem?

6. Is the problem such that pertinent data is made available? Are good data-collecting instruments available?

A problem is good and researchable if the researcher gets answers to the above questions in the affirmative.

WAYS IN WHICH A RESEARCH PROBLEM IS MANIFESTED

A problem is said to exist when we know that there is something we don't know really. There are at least three ways in which a problem is said to be manifested.

(1) A noticeable gap in the results of investigations

(2) Contradictory results of investigations

(3) Isolated fact in the form of unexplained information

These three may be discussed as given below.

(1) *A noticeable gap in the results of investigations*: A problem is manifested when there is a noticeable gap or absence of information. Suppose a community or group intends to provide psychotherapeutic services. Here, two questions arise: Firstly, what kind of psychotherapy they should offer and secondly, which one of the different forms of therapeutic methods is most effective for a given type of mental disease. These questions are highly important, but there are only a few scientific studies which provide answers to these questions. Therefore, a noticeable gap in the knowledge exists. Hence, there is the need for the collection of necessary data and their explanation for filling the gap.

(2) *Contradictory results of the investigations*: When several investigations done in the same field are not consistent and are therefore, at times, contradictory, a problem to find out a new answer and settle the controversy is easily manifested. If we review those experiments that have been conducted to examine the effect of reinforcement in learning, there are some experiments which confirm the fact that learning can't take place without reinforcement whereas there are some experiments which have vividly confirmed the fact that learning can take place even in the absence of reinforcement. Such a contradictory situation presents a problem. The investigators immediately want to know: Why does reinforcement affect learning in one situation but not in others? What are those situations in which reinforcement affects learning and those situations in which reinforcement does not affect learning? Likewise, several experiments have been published in which basically the same problem has been tried to be answered: What is the effect of the location of rest pause upon learning a task? Three independent and contradictory results have been obtained. One type of experiments has shown that progressively increasing rest periods are better for learning, another type of experiments has revealed that progressively decreasing rest periods are better for learning while the third type of experiments has revealed that effects of progressively increasing or progressively decreasing rest periods are about the same. Now, in such a situation the question arises: Why do these three types of experiments provide us with conflicting results? In an attempt to locate the possible reasons, the investigator may think of extraneous variables which might have been left uncontrolled and therefore, there were differences in the results of the experiment. So, the problem here arises: How to locate those extraneous variables, what their possible solutions can be, and so on.

(3) *Isolated fact in the form of unexplained information*: When the researcher possesses a fact and at the same time is interested in seeking the answer to the question, why this is so, he becomes aware of the problem. Putting it in other words, a fact which exists in isolation from the rest of our knowledge, demands an explanation. In fact, anything in itself presents no problem, but if it does not fit in with the existing knowledge, a problem is said to be manifested. In such a situation, collection of new information becomes necessary so that eventually the researcher can expect that the new fact will be related to the additional information in such a way that it will be adequately explained. This can be illustrated from Max Wertheimer's discovery of the phi-phenomenon which refers to the perception of a particular kind of movement in between stimuli that are actually stationary. It is said that Wertheimer, while sitting in a running train, saw movement from one telephone pole to another instead of seeing mere stationary telephone poles,

that is, first one and then another. Later on, Wertheimer systematically studied this phenomenon—using Kohler and Koffka, who later became important psychologists—as subjects. All three psychologists, that is, Wertheimer, Kohler and Koffka held that the phi-phenomenon could not satisfactorily be explained by the existing theories in psychology and, therefore, they advanced a new theory giving rise to what became known as Gestalt Theory in psychology. The essence of their viewpoints was that phi-phenomenon could not be understood by analyzing the situation into parts, rather, it could be understood only by studying the whole complex situation.

Thus, there are three primary ways in which the research problem is manifested. A research student must have, by now, adequate knowledge about these ways of manifestation.

TYPES OF RESEARCH PROBLEMS

Research problems can broadly be categorized into two types.

(a) Solvable problem

(b) Unsolvable problem

A solvable problem is one that raises a question which can be answered with the use of our normal capacities. An unsolvable problem is one that raises questions which can't be answered with our normal capacities. Unsolvable problems are usually concerned with supernatural phenomena. 'Does reward facilitate learning?' A problem like this is an example of a solvable problem. 'How does our mind work?', 'Is there a divinity that shapes our ends?'—these are examples of unsolvable problems. The chief distinction between a solvable problem and an unsolvable problem is that solvable problems can empirically be solved by studying observable events, whereas unsolvable problems can't be empirically studied.

Science addresses itself to only solvable problems. A problem is said to be solvable if it is possible to advance a suitable hypothesis as a tentative solution to it. A solvable problem, among other things, has the following two features:

(i) A solvable problem can be solved by empirical methods. An empirical method is one which depends upon observation of natural events.

(ii) In case of a solvable problem, it is possible to advance a suitable hypothesis as a tentative solution to it.

An unsolvable problem has three characteristics which help the researcher in recognizing it and making it distinct from a solvable problem. These characteristics are as follows:

(i) An unsolvable problem possesses the trait of unstructuredness. Such a problem remains unstructured because the interest of the researcher is not clear and the area it refers to is so unorganized that it no longer remains possible for the researcher either to specify the relevant observations or to relate the observations to these vague formulations. For example, 'Can human nature be changed?' is an example of such unsolvable problem in which sufficient degree of unstructuredness exists.

(ii) An unsolvable problem has inadequately defined terms and the operational definition of such a problem contains vagueness, and its variables are such that can't be defined operationally. In the above example, human nature is vague and it can't be defined operationally.

(iii) The third characteristic of an unsolvable problem is that relevant data can't be collected. Sometimes a problem is structured and its terms are operationally defined, yet the researcher can't collect relevant data.

These characteristics are such that an unsolvable problem can readily be distinguished from a solvable problem.

IMPORTANCE OF FORMULATING A RESEARCH PROBLEM

The formulation of a research problem is one of the first and significant steps in any research process. In fact, the formulation of a research problem is like the identification of a destination before starting a journey. As in absence of identification of destination, the researcher cannot start a journey by the correct route, so also in absence of formulating a good research problem, he cannot undertake a clear and defined steps of the research process.

A good research problem is like a good foundation of a building. As we know, the strength of a building depends upon the design and solid foundation, so also a good research and its outcome is dependent upon the formulation of a good research problem. Kerlinger (1986) has rightly opined that if the researcher really wants to solve a problem, he must know what the research problem is. He must have a clear idea with regard to what it is that he wants to find out and not only what he thinks he must find.

The research problem may be simple or complex. The way the researcher formulates the problem determines the steps to be followed in research: the type of variables to be included, the type of the research design that can be used, the type of sampling strategy to be employed, the tests/scales to be used as well as the type of analysis that can be undertaken. In simple words, it can be said that the formulation of problem is like the *input* into the research study and the *output* is the cause-and-effect relationship or association established. If the input is good, the output is bound to be good. On the other hand, if the input is vague and questionable, the output is bound to be questionable and unbelievable. Thus the famous saying about computers—*garbage in, garbage out* (GIGO)—is also applied to a research problem.

STEPS IN FORMULATING A RESEARCH PROBLEM

Formulating a research problem is one very important part of a research process. It consists of a number of steps, which a student or the researcher must know. These steps are described below.

1. *Identifying the broad field or area of interest:* The first step in formulating a research problem is to identify a broad field or subject that is of interest to the researcher. For example, if the researcher is interested in studying the treatment of abnormal behaviour of the person, he may conduct research on a subject area in the field of clinical psychology. If he is interested in how people form impression about others, he may select a research topic from social psychology. Likewise, if the researcher is interested in studying supervisory behaviour, he may undertake or explore the field of industrial/organizational psychology. These are the broad research areas, which are of interest to the researcher.

2. *Dissecting the broad area into subareas:* The next step in formulating a research problem is to dissect or divide the broad area into small subareas. For example, the field of clinical psychology may be dissected into small subareas like mental health of the persons, stressful behaviour, causative factors of mental retardation, types of phobic behaviour, etc. While the researcher is preparing the list of subareas, he should also have consultations with others who are experts and know the subject area well.

3. *Selecting any subarea that is of most interest to the researcher:* For any researcher, it is not feasible to study all subareas. Therefore, the researcher selects subareas that are of interest to him. Besides the interest of the researcher, there are many other considerations in selecting a problem and these have already been discussed. One easy way to select a subarea of interest is to decide through the process of elimination. The researcher goes through the list and deletes those subareas, which are of little interest. Through this process, gradually towards the end, he finds that he is left with the subarea, which is of high interest and which he considers manageable, and

that matches his level of expertise and related resources. Thus a subarea of the research is demarcated.

4. *Raising research questions:* After selecting the subarea, the researcher raises some questions. At this step, the researcher raises whatever questions he wants to answer in this subarea. If too many questions are raised, again through the process of elimination, he sticks to some manageable questions and eliminates other questions.

5. *Formulating objectives:* At this stage, the researcher formulates his main objectives and subobjectives of the research study. In fact, objectives grow out of the research questions. There is difference between objectives and research questions. Research questions are questions written in interrogative statements whereas objectives are something, which results from transformation of those questions by using, some action-oriented words such as 'to find out', 'to explore', 'to determine', 'to examine', etc. Thus objectives are derived from questions. Some researchers keep themselves satisfied with the research questions only and they don't formulate objectives. Besides this, there are researchers who proceed in reverse order, that is, they formulate objectives first and then, formulate the research questions.

6. *Assessing the objectives:* At this stage, the researcher examines the feasibility of achieving those objectives in light of his available financial and human resources as well as in terms of technical expertise available.

7. *Doing double-check:* This is a final step in formulating a research problem where the researcher goes back and rethinks whether or not he is really interested in the objective of the study and have sufficient and adequate step to undertake it. At this stage, the researcher usually asks questions like this: Do I have enough resources to undertake this study? Am I really interested and enthusiastic in undertaking this study? If the answer to any such question is 'no', the researcher tries to reassess his objectives.

Thus in formulating a research problem, a researcher has to go through these various steps sincerely and enthusiastically.

MEANING AND CHARACTERISTICS OF A GOOD HYPOTHESIS

In conducting research the next step after the selection of the problem is to formulate a hypothesis (pl. hypotheses). What is meant by a hypothesis? As we know, any scientific investigation starts with the statement of a solvable problem. When the problem has been stated, a tentative solution in the form of a testable proposition is offered by the investigator. This testable proposition is called a *hypothesis*. Therefore, a hypothesis is nothing but a suggested testable answer to a problem. Enlarging on this meaning of a hypothesis, we may say: A hypothesis is a testable relationship between two or more than two variables. Several experts have defined hypothesis in the same way. For example, the viewpoints of two leading experts can be quoted: McGuigan (1990, 370) has defined hypothesis as, "a testable statement of a potential relationship between two (or more) variables, that is advanced as potential solution to the problem." Kerlinger (1973, 18) has defined hypothesis as, "a conjectural statement of the relation between two or more variables. Hypotheses are always in declarative sentence form, and they relate, either generally or specifically, variables to variables." On the basis of these definitions, two points can be suggested about a hypothesis. First, a hypothesis is a testable statement, which means that it displays the relationship between those variables which are measurable or potentially measurable. Second, a hypothesis exhibits either a general or specific relationship between variables.

The above interpretation shows that the two general characteristics regarding hypotheses are identical to the two characteristics (already discussed) of problems. This should not produce the impression that problems and hypotheses are identical. In fact, there are some basic distinctions between a problem and a hypothesis. *First*, hypotheses are testable whereas problems are not. For example, "What is the relationship between IQ and classroom achievements?" is a problem which cannot be tested. However, when we deduce a hypothesis from this problem in the light of the available facts and theories, we can test it. Thus the hypothesis, "IQ is positively related with classroom achievement" can be tested directly. *Second*, problems are stated in the form of interrogative sentences whereas hypotheses are stated in the form of declarative sentences.

A hypothesis indicates what we are looking for. It is a proposition, which always looks forward. After testing the hypothesis, the investigator may find it to be correct or incorrect. If it proves to be correct, the problem (stated earlier) is solved and if it proves to be incorrect, the problem is not solved.

When a hypothesis has been formulated, the investigator must determine whether or not the formulated hypothesis is good. There are several *criteria* or *characteristics* of a good research hypothesis. A good hypothesis is one which meets such criteria or incorporates such characteristics to a large extent. Some of these characteristics are enumerated below.

1. *The hypothesis should be conceptually clear:* A good research hypothesis is one which is based upon operationally defined concepts. Not only this, the definitions must be given in commonly-accepted and communicable words so that the complete hypothesis is conceptually clear to any investigator.

2. *The hypothesis must be testable:* It should be formulated in a way that can be tested directly and found to be probably true or probably false. A hypothesis like, "Democratic atmosphere in an educational institution improves the creative thinking of the pupils," is an example of a hypothesis, which is too broad and hence, cannot be tested directly. Such a hypothesis, however, is potentially testable.

3. *The hypothesis should be economical and parsimonious:* If several hypotheses are offered to test a research problem, the more economical and parsimonious ones should be preferred to hypotheses involving higher monetary expenses.

4. *The hypothesis should be related to the existing body of theory and fact:* If the investigator advances a hypothesis, which seems to him of interest but which is not related to the existing body of theory of facts, it cannot be a good research hypothesis. If, for example, the investigator develops a hypothesis that skin colour produces a difference in intelligence then it may be an interesting hypothesis but may not be scientifically sound because no theory has proved so.

5. *The hypothesis should have logical unity and comprehensiveness:* If several hypotheses can be formulated regarding the same research problem, the most logical and comprehensive one should be preferred.

6. *The hypothesis should be general in scope:* A general hypothesis permits several deductions and thus, explains several facts at a time. Therefore, a general hypothesis should be preferred. However, very broad or very general hypotheses cannot be good research hypotheses because such hypotheses are often vague and cannot be tested.

7. *The hypothesis should be related to available scientific tools and techniques:* A hypothesis about which data cannot be collected because no scientific tools or techniques are available, cannot be a good research hypothesis.

8. *The hypothesis should be in accord with other hypotheses of the same field:* While this is not an essential condition, if any hypothesis satisfies this criterion, it can be claimed to be a good research hypothesis. A hypothesis that contradicts other hypotheses of the same field can also be regarded as a good hypothesis provided it is followed by scientific rationale, which in turn, has experimental support.

FORMULATING A HYPOTHESIS

It is difficult to tell precisely how a scientist formulates a hypothesis because the process of formulation itself is vague and idiosyncratic. Goode & Hatt (1952) have pointed out three major possible difficulties in formulation of a good research hypothesis. *First*, the absence of knowledge of a theoretical framework is a major difficulty in formulating a good research hypothesis. If detailed theoretical evidences are not available or if the investigator is not aware of the availability of those theoretical evidences, a research hypothesis cannot be formulated. *Second*, when the investigator lacks the ability to utilize the knowledge of the theoretical framework, a hypothesis cannot be formulated. *Third*, when the investigator is not aware of the important scientific research techniques, he will not be able to frame a good research hypothesis.

Despite these difficulties, the investigator attempts in his research to formulate a hypothesis. Usually the hypothesis is derived from the problem statement. The hypothesis should be formulated in a positive and substantive form before data is collected. In some cases additional hypotheses may be formulated after data has been collected, but they should be tested on a new set of data and not on the old set which has suggested it. The formulation of a hypothesis is a creative task and involves a lot of thinking, imagination and the like. Reichenbach (1938) has made a distinction between the two processes found commonly in any hypothesis-formulation task. One is the *context of discovery* and another is the *context of justification*. The manner or the process through which a scientist arrives at a hypothesis illustrates the context of discovery, and the presentation of evidence or proof in support of the truth of the hypothesis illustrates the context of justification. A scientist is concerned more with the context of justification in the development of a hypothesis. He never puts his ideas or thoughts as they nakedly occur in the formulation of a hypothesis. Rather, he logically reconstructs his ideas and thoughts and draws some justifiable inferences from those ideas and thoughts. He never cares to relate how he actually arrived at a hypothesis. He does not say, for example, that while he was shaving, this particular hypothesis occurred to him. He, usually, arrives at a hypothesis by the rational reconstruction of thoughts. When a scientist reconstructs his thoughts and communicates them in the form of a hypothesis to others, he uses the context of justification. When he arrives at a hypothesis, he extensively as well as intensively surveys a mass of data, abstracts them, tries to find out similarities among the abstracted data and finally makes a generalization or deduces a proposition in the form of a hypothesis.

For example, consider the following situation in Pavlovian conditioning. A bell is sounded and immediately after that, meat powder is presented to a dog. The dog starts salivating. After a number of presentations of the bell and food, the dog salivates at the mere presentation of the sound of the bell. In such a situation the experimenter observes that all the instances of salivation made in each trial are similar to each other and hence, they are abstracted by him as belonging to one general class of the salivation response. Likewise, the sound of the bell presented in each trial is seen by him as forming one general class because all the sounds of the bell are similar enough to form a class. Thus the scientist uses classification for distributing a mass of data into a smaller number of categories so that they can be effectively handled. Next, he tries to find out the relationship between the classified data so that a hypothesis can be formulated. For example, the experimenter may now formulate the hypothesis: After so many repetitions of the bell and food, the dog will eventually respond to the mere sound of the bell.

The above example illustrates the typical processes through which a scientist, at least a behavioural scientist, proceeds before arriving at a hypothesis. However, there is no dearth of scientists who proceed in a haphazard way in formulating a hypothesis.

WAYS OF STATING A HYPOTHESIS

A hypothesis is a testable statement of the relationship between the variables under study. Since it is testable, it can be shown whether a hypothesis is either true or false. When we are aware of these characteristics of a hypothesis, we must know how to state a hypothesis.

According to the suggestion given by L Russell (see Reichenbach 1947), the hypothesis should be stated in the logical form of the *general implication*. A hypothesis in form of a general implication may be expressed in terms of an "If... then...," relationship or "If *a* then *b*". In other words, if a condition holds true then some other conditions also hold true. For example, if the reward is given for learning a task then the learning is improved. In general implication, the *a* condition is referred to as the *antecedent condition* and the *b* condition is referred to as the *consequent condition*. The antecedent as well as the consequent condition is expressed in the form of propositions and therefore, the hypothesis expresses the relationship between the above two types of propositional statements. In stating the hypothesis as general implications, there occur two common misunderstandings regarding the antecedent condition and the consequent condition. *First*, it is generally misunderstood that the antecedent condition causes the consequent condition. This is not always true. The general implication only says that if the hypothesis is probably true then the antecedent condition and the consequent condition will occur together and not that one causes the other. *Second*, it is commonly misunderstood that the consequent conditions are always true. But in reality, the general implication does not say so. It only states that if the antecedent conditions are true then the consequent conditions are also true.

Now the question is: Are we following Russell's standard general implication in stating the hypothesis? The answer is probably 'no'. Recent publications of scientists have clearly shown that they are not following the "If the hypothesis in the form of a statement, which is direct and does not contain the "If ... then ..." form. For example:

(i) Reward improves learning.

(ii) The onset of fatigue reduces the efficiency of workers.

If we are not conforming to the advice of Russell, we are not necessarily committing a serious error in stating the hypothesis because the above type of hypothesis can be restated in the "If ..., then ..." form of Russell also. For example, "If a reward is given, learning is improved" and "if fatigue occurs, the efficiency of the worker is reduced."

It is obvious that the hypothesis can be formulated in different ways.

TYPES OF HYPOTHESES

On the basis of the degree of generality, research hypothesis can be divided into two types.

(a) Universal hypothesis

(b) Existential hypothesis

A universal hypothesis is one in which the stated relationship holds good for all the levels or values of variables which are specified for all time at all places. 'Adequate level of light increases reading efficiency' is an example of universal hypothesis.

Existential hypothesis is one which states that the relationship stated holds good for at least one particular case. For example, 'There is at least one schizophrenic who does not have either delusion or hallucination' is an example of existential hypothesis.

Of these two types of hypotheses, the universal hypothesis is preferred because such a hypothesis has a greater predicative power than the existential hypothesis.

As we know, hypothesis is a formally stated expectation about a behaviour that defines the purpose and goals of the study being conducted. Based upon the goals of explaining and controlling the causes of behaviour, there are two types of hypothesis.

1. Causal hypothesis
2. Descriptive hypothesis

1. *Causal hypothesis*: A causal hypothesis postulates a particular causal influence on behaviour. In other words, it tentatively explains a particular influence on, or a cause for, a particular behaviour. For example, if the researcher hypothesizes that boring contents of commercial advertisements is the cause of channel changing by TV viewers it becomes the example of causal hypothesis. Although it is a fact that boring contents may not be the only causal influence of the channel-changing behaviour, it is the probable cause we are investigating at the moment.

2. *Descriptive hypothesis*: Descriptive hypothesis is one that postulates particular characteristics of a behaviour or provides some specific goal for the observation. In fact, such hypothesis tentatively describes a behaviour in terms of the characteristics or the situation in which it occurs. Such hypothesis identifies the various characteristics or attributes of behaviour and allows us to predict when it occurs. For example, the researcher hypothesizes that channel changing during TV viewing occurs more frequently when the person is alone than when he is watching with others. The reality may be that even the number of people present might partially cause channel changing, and the researcher has not stated that. In this way, it can be said that a descriptive hypothesis simply describes the behaviour in terms of the various characteristics of the situation and it does not attempt to identify the causes of a behaviour.

Apart from these, the other type of hypotheses that we commonly use in behavioural researches are *simple hypothesis, complex hypothesis, research hypothesis, alternative hypothesis, null hypothesis* and *statistical hypothesis*. These may be described as under.

1. *Simple hypothesis*: Simple hypothesis contains only one or two variables. For example, hypotheses like children from broken homes tend to become delinquent, reward improves learning, aggression is associated with frustration are all examples of simple hypotheses. In all these hypotheses, the relationship between only two variables have been postulated. Hence, they are examples of simple hypotheses.

2. *Complex hypothesis*: Complex hypotheses are hypotheses which contain more than two variables and therefore, require complex statistical calculation too. Such hypotheses are called complex because the interrelatedness of more than two variables acting simultaneously is more difficult to assess quantitatively and theoretically. A hypothesis like children from upper and lower socio-economic status have larger adult adjustment problems than children from middle socio-economic status is an example of a relatively complex hypotheses.

3. *Research hypothesis*: A hypothesis derived from the researcher's theory about some aspects of behaviour is called a *research hypothesis* or is also known as a *working hypothesis*. The researcher believes that his research hypotheses are true or that they are accurate statements about the conditions of things he is investigating. He also believes that these hypotheses are true to the extent that the theory from which they were derived is adequate. In this perspective, Siegel and Castellan (1988) have defined research hypothesis as, "the prediction derived from the theory under test."

4. *Alternative hypothesis:* The formulation of an alternative hypothesis (H_1) is a convention in scientific research. Alternative hypothesis is defined as the operational statement of the investigator's research hypothesis. It functions as an alternative to null hypothesis and typically asserts that the independent variable has an influence upon dependent variable. If null hypothesis is rejected by the results of the experiment, alternative hypothesis is accepted and the research hypothesis is strengthened. Alternative hypothesis is sometimes also called as *experimental hypothesis* or *statistical hypothesis*.

5. *Null hypothesis*: A null hypothesis (H_0) is, in a sense, the reverse of a research hypothesis. It is, in fact, a no-effect or difference hypothesis or negation hypothesis that tends to refuse or deny what is explicitly indicated in a given research hypothesis. Generally, the experimenter or researcher's aim is to refute this hypothesis on the basis of the obtained results so that its reverse, that is, the research hypothesis can be supported or confirmed.

The interrelatedness of research hypothesis, null hypothesis and alternative, or statistical hypothesis can be explained through an example: Suppose a certain social–psychological theory would lead us to predict that two specified groups of people would differ on the measure of intelligence. This prediction would be our research hypothesis which would state that the two groups differ. Confirmation of this hypothesis would lend support to the theory from which it was derived. To test this research hypothesis, we state it in an operational form as the alternative hypothesis, that is, H_1. One operational way to state this alternative hypothesis would be that mean intelligence scores of these two groups differ or are unequal. The null hypothesis (H_0) would be that the mean intelligence score of the two groups is the same. If the data permits us to reject H_0 then H_1 would be accepted because the data support the research hypothesis and its underlying theory.

In fact, the nature of the research hypothesis determines how the alternative hypothesis (H_1) should be stated. If the research hypothesis simply states that two groups will differ with respect to mean then the alternative hypothesis would be simply that mean of the two groups are not equal. But if the research hypothesis predicts *difference with direction*, that is, one specified group will have a larger mean than the other then the alternative hypothesis may be that the mean of group 1 is greater or less than the mean for group 2.

SOURCES OF HYPOTHESES

In social science researches, a researcher generates a hypothesis from several sources. Some of these important sources are as follows:

1. One primary source of generating a hypothesis is various opinions, observations and experiences. A researcher bases his hypothesis on these sources so long as he conducts an empirical and objective study to provide evidence for testing the hypothesis.

2. Another source of a hypothesis is retesting of a hypothesis previously tested by some other researcher.

3. A third source of generating a hypothesis is the existing research itself. A talented researcher, while reading about the results of a reported study which tested one or more than one hypotheses, often generates several additional hypotheses. For example, if the researcher finds that positive reinforcement improves learning, he would then like to determine why so and may also try to identify the factors that modify this influence.

4. A fourth source of a hypothesis is *theory*. A theory, as we know, is a logically organized set of such proposals that defines, explains and organizes our knowledge about many behaviours. In fact, a theory involves abstract concept that applies to a broad range of behaviours in a parsimonious way. For example, Freudian theory tried to explain and relate a vast array of normal and abnormal behaviours using several abstract concepts like ego, superego, unconscious mind, subconscious mind, etc. Theoreticians generally proceed in either of the two ways: *First*, they may develop a theory beginning mainly with some ideas and concepts for which there is little scientific evidence. In such situations, a theory provides a direction for the evidence that the researchers will seek. *Second*, theoreticians may develop a theory after substantial evidence has been collected. In such situations, a theory provides a way for organizing diverse findings. In either of these two situations, researchers derive a specific hypothesis, test it by conducting a study, and then, apply the results of the study back to the diversified theoretical concepts, either adding or modifying the theory. Then, from this modified new theory, the researcher further generates some additional hypotheses.

5. The fifth source of a hypothesis is a *model* which is defined as a generalized hypothetical description that, by analogy, explain the various processes underlying a set of common behaviours (Heiman 1995). The basic difference between a theory and a model is that a theory accounts for broad, abstract components of behaviours whereas a model provides a more concrete analogy for thinking about the various concepts and how they actually operate. The basic characteristic of a psychological model is that it involves a flow chart or diagram of a certain psychological process. For example, 'Information processing model' describes human memory with help of a flow chart having three separate boxes that represent the flow of information through short-term and long-term memory (Atkinson & Shiffrin 1968). This explanation of the model provides a useful analogy for deriving specific hypotheses about how and when information remains in memory for the time being or more permanently. Subsequently, these hypotheses are tested and the obtained results are used to bring a modification in the model, if needed.

In this way we find that there are several sources from which the researcher derives hypotheses for his study.

FUNCTIONS OF HYPOTHESES

For a behavioural scientist, a hypothesis does the following major functions.

1. *Hypotheses test various theories*: In behavioural research the researcher develops a theory to account for some phenomenon and then, he devises a means whereby the theory can be tested. He seldom tests the theory directly. Most of the time he conducts tests of hypotheses that have been generated and derived from that theory. If the hypotheses test out as specified by the researcher, it is said that his theory is supported in part. Thus one of the major functions of hypotheses is to make possible to test theories. Thus in this light, an alternative definition of a hypothesis can be the statement of theory in a testable form. All statements of theory in testable form can be called hypotheses.

2. *Hypotheses suggest various theories*: In behavioural research, it is often found that some hypotheses are not associated with any particular theory. It is just possible that as a result of some hypothesis, a theory may eventually be constructed. Therefore, another function of hypotheses is to suggest theories that may account for some event.

Although it is a common practice that the researcher proceeds from theory to hypotheses, occasionally the reverse is true. The researcher may have some idea about why a given phenomenon occurs and he may hypothesize a number of things that relate to it. He may find that some hypotheses have greater potential than others for explaining the event or particular behaviour, and as a result, he may construct a logical system of propositions, assumptions and definitions linking his explanation to the event. In other words, it can be said that he has devised a theory.

3. *Hypotheses tend to describe social phenomena*: A hypothesis also does a descriptive function. When a researcher tests a hypothesis empirically, it tells him something about the phenomenon it is associated with. If the hypothesis is supported then his information about the phenomenon increases. When the hypotheses are refuted, the test tells us something about the phenomenon the researcher did not know before. The accumulation of information as a result of the hypothesis testing that way reduces the amount of ignorance of the researchers.

Apart from these primary functions, hypotheses also have some important secondary functions. In fact, testing hypotheses tend to refute some 'common sense' notion about the behaviour of the organism, raise questions about the explanations we presently use to account for things and also tend to alter our orientation towards the environment to one degree or another. Besides these, as a result of testing certain hypotheses, social policy may be formulated in communities, delinquents and offenders may be treated differently, teaching methods may be modified and improved, solutions to various kinds of problems may be suggested and penal institutions may be redesigned and revamped. Besides, hypothesis also tells the researcher what data to collect and what not to collect, thereby providing focus to the research study.

☑ Review Questions

1. What is a hypothesis? Point out the characteristics of a good research hypothesis.
2. Explain the significance of hypothesis in research. What are the main considerations in formulation of a hypothesis? State them.
3. What is the difference between a research problem and a research hypothesis? State the sources of research problem.
4. Discuss the major steps in formulating a research problem.
5. Discuss the main sources of formulation of a hypothesis. Also, point out the importance of different types of hypotheses in behavioural researches.
6. Citing relevant examples, discuss the ways in which a research problem is manifested.
7. Write short notes:
   (a) Types of problems
   (b) Types of hypotheses
   (c) Important considerations in selecting a problem

◆

19

REVIEWING THE LITERATURE

CHAPTER PREVIEW

- Purpose of the Review
- Types of Literature Review
- Sources of the Review
  - Journals and Books
  - Reviews
  - Abstracts
  - Indexes
  - Internet
  - Doctoral Dissertations
  - Supervisors/Research Professors
- Types of Literature
- Writing Process of the Literature Review
- How Old should the Literature Be?
- Preparation of Index Card for Reviewing and Abstracting
- Abstract

PURPOSE OF THE REVIEW

A collective body of works done by earlier scientists is technically called the *literature*. Any scientific investigation starts with a review of the literature. In fact, working with the literature is an essential part of the research process which generates the idea, helps in developing significant questions and is regarded as instrumental in the process of research design. The main objectives of a review of the literature are enumerated below.

1. *Identifying variables relevant for research* : When the researcher makes a careful review of the literature, he becomes aware of the important and unimportant variables in the concerned area of research. A careful review also helps the researcher in selecting the variables lying within the scope of his interest, in defining and operationalizing variables and in identifying variables which are conceptually and practically important. Thus a review of the literature, on the whole, prepares the researcher to formulate a researchable problem in which conceptually and practically important variables are selected.

2. *Avoidance of repetition*: A review of the literature helps the researcher in avoiding any duplication of work done earlier. A careful review always aims at interpreting prior studies and indicating their usefulness for the study to be undertaken. Thus prior studies serve as the foundation for present studies. In some cases, the duplication or replication of prior studies becomes essential. This is especially true when the researcher wants to test the validity of the earlier studies. In such a situation, too, a careful review helps the researcher in getting acquainted with the number and nature of the studies related to the study whose validity is being assessed at present.

3. *Synthesis of prior works*: A careful review of the literature enables the researcher to collect and synthesize prior studies related to the present study. This, in turn, helps him in building a better perspective for future research. A synthesized collection of prior studies also helps a researcher to identify the significant overlaps and gaps among the prior works.

4. *Determining meaning and relationship among variables*: A careful review of the literature enables the researcher in discovering important variables relevant to the area of the present research. When significant variables are discovered, the relationship among them can be identified. Subsequently, the identified relationship is incorporated into different hypotheses. Thus, for conducting a scientific study, the relationship between the different variables must be explored by reviewing the literature so that a good context may be built up for subsequent investigations.

In addition to these specific purposes, there are some general purposes of the literature review.

(i) To argue for the relevance and the significance of the research question

(ii) To provide the context for one's own methodological approach

(iii) To establish one's own credibility as a knowledgeable and capable researcher

(iv) To argue for the relevance and appropriateness of one's own approach

TYPES OF LITERATURE REVIEW

As we know, the basic aim of review of literature is to integrate and summarize what is known in the area of interest. There are many types of literature review. Some of the common and popular types are given below.

(a) *Historical review:* This is a type of research review in which the researcher traces an issue over time. Such review is sometimes merged with a theoretical or methodological review to demonstrate how concept, theory or research method had developed over time.

(b) *Context review*: Here the researcher tries to link a specific study to the larger body of knowledge. Such review often appears at the beginning of a research report and introduces the study within a broader framework and thus it continues to build on developing a line of thought.

(c) *Integrative review:* It is a very common type of review where the researcher presents and summarizes the existing state of knowledge regarding a topic or theme, and highlights agreements and disagreements within it. Sometimes it is combined with context review or may be published as an independent article as a probable service to new researchers.

(d) *Methodological review:* It is a specialized type of integrative review where the researcher tends to compare and evaluate relative methodological strength of various studies and show different methodologies such as research design, samples, etc., for different type of results. Meta-analysis is often used in methodological review.

(e) *Self-study review:* A type of review where the researcher displays his or her familiarity with the topic of interest. It is often part of an educational programme.

(f) *Theoretical review:* It is a type of specialized review where the researcher presents several concepts, models, theories, which concentrate on the same topics and compares all of them on the basis of assumptions, scope of explanation, logical consistency, etc.

Of these various types of review of literature, integrative review, context review and theoretical reviews are comparatively popular.

SOURCES OF THE REVIEW

There are different sources of the review of literature. Some of them are enumerated below.

Journals And Books

Different research journals and books relevant to the areas of interest are the primary sources of the literature review. Most major libraries have a periodical section where different types of research journals are made available easily. A research journal generally contains the publication of original research reports with their detailed methodology and results. Such journals are *referred* and therefore are different from *nonreferred* journals. A referred journal is one which reports only those articles which are carefully reviewed by the experts before publication. Often, the reviewer rejects several manuscripts and selects a few for publication. Similarly, books are also direct sources of the literature review and they include textbooks, handbooks, yearbooks, dictionaries, etc. Of these two, journals are regarded as more useful because they provide the researcher with the latest and up-to-date information relevant to the area of interest.

Reviews

Reviews are short articles that give brief information regarding the work done in a particular area over a period of time. Reviews are commonly published in *journals, yearbooks, handbooks* and *encyclopaedias*. Reviewers select research articles of their interest, organize them contentwise, criticise their findings and offer their own suggestions and conclusions. Review articles are a good source for those investigators who wish to have all the relevant researches at one place without taking pains to look for them. Since the reviewers organize all the possible research papers of the relevant area in their review articles, review articles also provide the advantage of prior reviews.

Abstracts

Abstracts provide a summary of the research reports done in different fields. *Psychological Abstracts* (Washington: American Psychological Association) first published in 1927, and *Sociological Abstract* (New York: Sociological Abstracts, INC) are the two common examples of abstracts. Other sources of abstracts are Educational Resources Information Centre (ERIC), Dissertation Abstracts International, children research abstract and the Indian Education Abstracts (published by NCERT, New Delhi). These abstracts are useful sources of up-to-date information for researchers. In an abstract, besides a summary, researchers get all the relevant information such as the title of the research report, name of the author and the journal pagination information, etc., regarding the research article. The only limitation of abstracts is that they fail to satisfy those researchers who desire detailed information regarding the methodology and results of the research articles.

Indexes

Indexes show the titles of the research report without any abstract. The titles are categorized and arranged alphabetically in each category so that the researcher can locate any article of interest easily. The *Education Index* (New York: H W Wilson Co.) is a good example of an index. As indexes do not provide detailed information, they keep many a researcher dissatisfied. They can be best regarded as a supplementary source which, if combined with other sources, can yield valuable information to the researchers.

Internet

Today the internet is a very easy and quick source of literature review. Internet sites are very useful for providing easy access to original writings by important researchers. They also provide an updated information on the topic that ordinarily is not available in the library. Internet sites also provide useful bibliographies related to a particular researcher. Search on the internet also reveals some relevant professional societies and academic associations which can provide a lot of support to the studies in the concerned area. Such organisations also sometimes publish important papers or periodicals which can be of immense help to the researchers. Some publishers put the brief content and extracts from recently published books on the Internet and these can be of valuable help to the researchers. Sometimes, the internet sites include articles extracted from encyclopedias which can also be very useful and informative as background reading. However, they are not normally suitable for citing in a thesis.

Doctoral Dissertations

Doctoral dissertations have also been a very good source of the review of the literature. In libraries of universities, doctoral dissertations are available. The researcher can choose the dissertations of their interest and find useful and relevant information there. There are no set forms of writing the research report in a doctoral dissertation but most dissertations contain chapters like an introduction, review of the literature, purpose of the study, method of the study, results, discussion, summary and conclusion. Some researchers prefer not to add a separate chapter on the review of the literature and hence, it is incorporated into the introduction itself. Thus the doctoral dissertations present the advantage of prior review. Ordinarily, it is not possible for the researcher to move through all the important libraries in the country to consult all existing doctoral dissertations. Hence, he can have access to those dissertations that interest him through *Dissertation Abstracts International*, which publishes the abstracts of the doctoral dissertations submitted to different universities. In India, the *Survey of Research in Education* (edited by M B Buch) does much the same function. The second *Survey of Research in Education* covering the period between 1972–78 has been released. Recently, the listing of dissertation abstracts has been computerized through DATRIX in terms of the key words (usually words appearing in the title of the dissertation).

Supervisors/Research Professors

Supervisors often know the literature well and are able to guide researches in the right direction. They are the recognised authority on the topic or research problems. Therefore, they should be consulted and their suggestions and advice should be carefully analyzed. It may also be that other research professors have recently sourced and reviewed the literature or an area very close to the literature that the researcher is seeking. So they also constitute one important source.

Whatever may be the sources of reviews, the process of reviewing literature itself is not above criticisms. Inevitably the interpretation of findings, insights derived, the manner in which conclusions are drawn are all solely dependent upon the judgements of the reviewer. In other words, such reviews fall prey to what is called *subjective judgement*.

TYPES OF LITERATURE

In order to work with appropriate literature, it is essential that the researcher must be able to identify and find it. For this, he must have an understanding of various literature types. Some of the common types of literature are given below.

Subject-specific books

Introductory and advanced textbooks and research reports can provide important background and context for the research. Such literature also provide information about theory and method of the research.

Grey literature

By grey literature is meant both published and unpublished materials that somehow don't have International Standard Book Number (ISBN) or an International Standard Serial Number (ISSN).

Grey literature is a broad category that includes unpublished research theses, newspaper articles, conference papers and pamphlets. During the course of the research, most researchers generally utilize one or the other type of grey literature.

Official publications, archives and statistics

This type of literature serves a dual purpose. Firstly such literature can be a valuable source of background and contextual information and secondly, they can also be used as a source of secondary data. Document and secondary data analysis are often based upon this type of literature.

Writing aids

As its name implies, such literature generally offer a significant support during the process of writing and can be easily used to improve the linguistic style of the work. Such literature includes dictionaries, bibliographic works, encyclopedias, thesauruses, yearbooks, books of quotes, almanacs, etc.

Journal articles

This type of literature is very common among the researchers. Its popularity is due to several factors. *First,* journal articles are very credible. *Second,* they are often targeted for academic audience. *Third,* they possess the trait of speciality. *Fourth,* they possess the regularity of production which means that research articles are not only relevant but also current.

Of these various types of literature, the relevant one is searched by the researcher.

Although reviewing literature is time-consuming and daunting, it is also a reward otherwise because it serves several functions. Of these functions, the following ones are important:

(i) It provides a strong theoretical background to the study.

(ii) It helps the researcher in contextualising the findings of the study.

(iii) Reviewing literature helps the researcher to show how his findings have contributed to the existing body of knowledge.

(iv) Reviewing literature also helps the researcher in establishing a link between what he is proposing and what has already been studied. In simple words, it helps the researcher in refining the research methodology.

(v) Reviewing literature also helps the researcher in clarifying and focusing on the research problem.

WRITING PROCESS OF THE LITERATURE REVIEW

Since the literature review may be a very long chapter, it does need some form of structure. The simplest ways of organizing the research works is to discuss them in chronological order. But this way may not prove to be appropriate in all situations. Another way may be that works on different subjects may be grouped together with the date of publication as only criterion of order. But this may also be confusing. Still another way may be to base the structure on the different types of publications. For example, chapters from books, journal articles and books by single authors should be separately grouped and structured. The basic aim of literature review is to use the literature for *informing, establishing* and *arguing*. In fact, literature review should go beyond a *he said/she said report.* O'Heary (2004) has recommended that for writing a good literature review, the following steps should strictly be followed:

(i) *To read a few good, relevant reviews* Researcher should take a look on the literature reviews done in several theses and journal articles. From these reviews, good and relevant reviews should be sorted and this depends upon the research skills of the researchers. The supervisor should help him in selecting the relevant and good reviews.

(ii) *To write critical annotations while going through various reviews* The researcher should begin to sort and organize the annotations of the reviews by themes, issues of concern and common limitations, etc. While he will be doing so, some patterns would start emerging and this would, in turn, help him in developing his own argument.

(iii) *To develop a structure* The researcher should structure the potential reviews according to his most urgent needs such as topical themes, arguments he wishes to establish, etc. The structure so developed is always subject to modification with the emergence of new thinking.

(iv) *To write purposefully* The researcher should note that he can review the literature without any agenda but he cannot write a formal literature review without any definite agenda or aim. The readers must know the reasons why and what the researcher is telling them.

(v) *To use the literature to support the argument* The researcher should not use the review only for reporting or borrowing the arguments from others, rather he should use the literature for generating ideas that may help or support his own arguments.

(vi) *To make the literature review an ongoing process* The researcher should make the literature review an ongoing process. In other words, his literature review should inform the researcher's questions, theories and methods and these should help in setting the parameters of the literature review. Thus literature review becomes a cyclical process and should often have a moving target.

(vii) *To get plenty of feedback* The researcher should not wait up to the last minute of writing process. Whatever has been written should be passed over to supervisors and other experts for their feedback. Early feedback gives a chance for rethinking and modification of ideas which can be incorporated in the writing process.

(viii) *To remain prepared for redrafting* In view of the suggestions through feedback, the researcher should redraft the review in a coherent manner so that his argument is reasonably supported.

Thus writing the literature review is a complex task which can be made easy by following certain steps.

HOW OLD SHOULD THE LITERATURE BE?

One of the most important questions before the researcher is: how old may the literature be? The simplest answer to this question is that it can be of any age. In fact, academic research is a cumulative activity. Each generation of researcher learns from the work of previous generation and current research basically depends upon the work and insights of the previous researchers. Since in any society the latest and contemporary research and publications are subject to vigorous demand and due to this reason, it is preferable to cite recent publications as far as possible.

Despite this, almost in any discipline, there are some seminal works which are centuries old but have become so significant that they are still being preferred by the researchers. Although their original ideas have been modified by subsequent researchers over time, their original spirit and views still remain significant and are held in considerable esteem. For example, the work of Sigmund Freud in the field of psychoanalysis was done about 150 year back but his ideas, theories, viewpoints are so pertinent and of importance that any researcher of today, working in this field, is bound to originate the review of the literature from his valuable contributions.

However, it would be a healthy suggestion for researchers that they should always take precaution in citing older works unless they are confident and convinced in quoting them.

PREPARATION OF INDEX CARD FOR REVIEWING AND ABSTRACTING

After going through the different sources of the review of the literature, researchers prepare their own review and abstract on the index card. Usually for this purpose, a 6" x 10" index card is recommended. In most journal articles, an abstract in about 150 words is provided. The researcher can incorporate it in his own abstract. Where the article seems to be very important and relevant, the researcher can prepare a more detailed version. Usually, the abstract, thus prepared, is divided into three parts. The *first* part consists of the *purpose* and *hypothesis* of the study. The researcher should write down the purpose of the study in not more than two lines. If the hypotheses are small, they can be recorded verbatim but if they are lengthy, they should be synthesized. The *second* part consists of the *methodology* of the study in which the size of the sample, nature of the population, methods for measuring or manipulating the variables, methods of data collection, designs and statistics are shown in synthesized form. The *third* part consists of the *findings* and *conclusion*. In this part, the researcher should briefly write down the findings relating to each hypothesis and also the conclusion drawn by the author concisely. At the top of the index card, a full reference should be clearly written in exactly the same way in which it appears in the researcher's own reference list. There are different types of research formats (which will be discussed in Chapter 24) but that followed by the *Publication Manual of the American Psychological Association* is widely popular and has been adopted by most of the important research journals. The researcher should never trust his memory for recall of the details of any research article and therefore, all the important and relevant details should be carefully noted down in the index card. A sample of the review abstract adjusted to the length of the 6" × 10" index card has been given below.

ABSTRACT

Singh, A K (1980). Extraexperimental Associations in Proactive Interference. *Journal of Psychological Researches*, 24 (3): 133–39.

Purpose And Hypotheses

This study explores the possibility of extraexperimental associations as one source of proactive interference (PI). There were four hypotheses: (1) US (trigrams having uncommon sequences) produces high amount of PI as compared to CS (trigrams having common sequences) and NS (trigrams having neutral sequences); (2) NS produces higher amount of PI than CS; (3) CW (words for middle frequency) produces higher amount of PI than NW (words of middle frequency) and UW (words of low frequency); (4) NW produces higher amount of PI than UW.

Methodology

In this study 30 subjects (all college students) were required to memorize six experimental lists to the criteria of one perfect recall by the method of serial anticipation. Subsequently, 30 subjects were divided into three groups of 10 each. A recall test was taken immediately after 2 minutes for Gr. I, after 2 hours for Gr. II and after 2 days for Gr. III. The amount of PI occurring over a 2-minute interval (Gr. I) served the main baseline for comparison. The significance of the mean differences of recall by the three groups was tested through the t test.

Findings And Conclusion

US produced the highest amount of PI and was followed by NS. Among words, CW produced the highest amount of PI whereas UW produced the least amount of PI. The amount of PI in case of NW was intermediate. It was concluded that linguistic associations (both meaningful and nonsense) in subjects before entering the laboratory tend to increase the amount of PI.

Thus the reviewing and abstracting of the literature on the index card should be done carefully and systematically. Sometimes it has been reported that researchers trust their memory for recalling a particular detail. But this is not a healthy practice because they are apt to forget the details or their memory may be blurred after some time. Researchers should try their best to accommodate every important and relevant detail under the three common headings suggested above.

☑ Review Questions

1. What is meant by the 'review of literature'? Discuss its major purposes.
2. Discuss the major sources of review that are commonly undertaken by the researcher in the field.
3. Discuss the various types of literature. How old should the literature be?
4. Point out the steps involved in writing a good literature review.
5. Discuss the different types of literature review.

◆

20

VARIABLES

CHAPTER PREVIEW

- Meaning and Types of Variables
    - Dependent Variables, Independent Variables and Intervening Variables
    - Qualitative Variables and Quantitative Variables
    - Continuous Variables and Discrete Variables
    - Moderator Variables and Intervening Variables
    - Active Variables and Attribute Variables
- Difference between a Variable and a Concept
- Methods of Measuring Dependent Variables
- Important Considerations in Selection of Variables
- Important Approaches to Manipulating Independent Variables
- Techniques of Controlling Extraneous Variables
    - Technique of Elimination
    - Constancy of Conditions
    - Balancing
    - Counterbalancing
    - Randomization
- Controlling Demand Characteristics

MEANING AND TYPES OF VARIABLES

A variable, as the name implies, is something which varies. This is the simplest and the broadest way of defining a variable. However, a behavioural scientist attempts to define a variable more precisely and specifically. From his point of view, variables may be defined as those attributes of objects, events, things and beings, which can be measured. In other words, variables are the characteristics or conditions that are manipulated, controlled or observed by the experimenter. Intelligence, anxiety, aptitude, income, education, authoritarianism, achievement, etc., are examples of variables commonly employed in psychology, sociology and education.

Variables have been classified in three different ways.

(i) From the viewpoint of causation
(ii) From the viewpoint of the design of study
(iii) From the viewpoint of unit of measurement

A discussion follows.

Dependent Variables, Independent Variables and Intervening Variables

From the viewpoint of causation or attempt to investigate a causal relationship or association, four sets of variables usually operate in the study.

- *Change variables,* which are responsible for bringing change in the phenomena.
- *Outcome variable,* which are the outcome or effect of the change variable.
- *Unmeasured variables,* that affect the link between cause-and-effect variable.

- *Connecting Variables,* which in certain situations are considered essential for completing the relationship between cause-and-effect variables.

In the terminology of research methodology, the change variables are called as *independent variables,* the outcome variables are called *dependent variables,* the unmeasured variables that affect the link between cause-and-effect relationship are called as *extraneous variable* and the variables that link a cause-and-effect relationship are called *intervening variable* or sometimes *confounding variables* (Grinnell 1993).

The terms 'dependent variable' and 'independent variable' have been borrowed from the field of mathematics in behavioural researches. The classification of variables into dependent and independent is frequently employed in experimental research. The dependent variable (DV) is defined as one about which the experimenter makes a *prediction*. The independent variable (IV) is defined as one which is *manipulated, measured* and *selected* by the experimenter for the purpose of producing observable changes in the behavioural measure (or DV). In other words, the independent variable is the variable on the basis of which the prediction about the DV is made. The occasional synonym of IV is *controlled variable,* which is rarely used because of its confusing nature with control variable. Underwood (1966, 12) calls the IV as the *stimulus variables* and the DV as the *response variables.* An example may illustrate the distinction between the IV and the DV. Suppose the experimenter wants to study the effect of teaching methods upon the classroom achievement of pupils. For this purpose, he may employ three methods of teaching, say, A, B and C and may teach the same group of pupils by these three methods and subsequently, the achievement may be measured or predicted. Teaching methods constitute the example of the IV and the classroom achievement constitutes the example of the DV. Similarly, if the experimenter wants to study the effect of a religious group upon attitude towards family planning, he may take the Hindu, the Muslim, the Sikh and the Parsi as the four religious groups and study their attitudes towards family planning. Subsequently, he may be able to predict which religious group has a favourable attitude or unfavourable attitude towards family planning. In this example, the religious groups constitute the example of the IV and the attitude towards family planning constitutes the example of the DV.

Sometimes in the experiment, a variable is left uncontrolled and is unintentionally allowed to vary with independent variable. Such a variable is called a *confounding variable*. For example, suppose the researcher is to assess the effectiveness of the lecture method vs. laboratory demonstration method in acquiring knowledge of fundamentals of statistics. There may be two equivalent groups of students. One group is lectured by Professor A and the other group with laboratory demonstration is taught by Professor B. Subsequently, both groups receive the same final examination and their performances are compared. Statistical analysis reveals that students taught by Professor B excel in their performance as compared to the students taught by Professor A. Can the researcher conclude that method of laboratory demonstration is better than the method of lecture? The answer is emphatically *no*. The method of instruction is not the only way the groups differ in terms of how they have been treated. The instructor (Professor) was allowed to vary along with the method of instruction. It might be that Professor B is a better instructor, motivates students properly and can explain concepts in understandable terms. On the othe hand, Professor A might be less experienced. That is, the difference between groups i performance might be due to the different instructors rather than the different treatments. It i difficult to be sure which interpretation is correct. This is an example of *confounding variable* an when a study is confounded, it is impossible to arrive at meaningful conclusions about th results.

One important thing related to the selection and measurement of dependent variable is th of floor and ceiling effect. A *floor effect* is said to have occurred where a null result occu because majority of subjects score at the very bottom end of the scale. Such an effect usual emerges when the task is too difficult for the subjects. For example, if in any experiment t purpose is to compare people's sensitivity to lights of different colours but all the given colours are below thresholds, no meaningful data will be obtained. This illustrates floor effect. Ceiling *effects* are the opposite of floor effects and are said to have occurred when subjects score too close to the top of the scale simply because the task was too easy. Continuing with the same example, if all the coloured lights are clearly visible because they are bright, the task would be too easy for the subjects and therefore, ceiling effect will occur. These two types of effects can be obtained with any task whose dependent variables cannot track the full range of independent variable. These effects must be prevented and this can be readily done through pilot study that can check the appropriateness of the subject pool and variables before carrying out the experiment proper.

In addition, a dependent variable must be sensitive, valid and reliable. To be sensitive, it must change clearly when the independent variable changes. To be valid, it must measure what it is intended to measure. To be reliable, a dependent variable must act consistently in relation to the independent variable. If subjects respond a certain way on one occasion but in a wholly different way on another occasion when the independent variables remain the same, the dependent variable would be unreliable.

As it has been said above, the IV is manipulated by the experimenter and its effect is examined upon the DV. Some experts, depending upon the mode of manipulation, have tried to divide the IV into *Type-E independent variable* and *Type-S independent variable* (D'Amato 1970, 11). Type-E independent variable is one which is *directly* or *experimentally* manipulated by the experimenter, and Type-S independent variable is one which is manipulated through the process of selection only. Such variables are difficult to manipulate experimentally or directly. An example may illustrate the distinction between the Type-E and Type-S independent variable. Suppose the experimenter wants to study the effect of temperature upon the rate of production in an industry. Here the IV is the temperature and the DV is the rate of production. He may manipulate the temperature by dividing it into three categories—high, medium and low—and examine its effect upon the rate of production. Here the temperature is being directly manipulated by the experimenter and hence, it constitutes the example of Type-E independent variable. Suppose, for the time being, that the experimenter is interested in answering the question: Is the rate of production dependent upon the age of the workers? Age is here the independent variable. For investigating this problem, the experimenter will have to select groups of workers on the basis of their ages in a way by which he can get an appropriate representation from different age groups ranging from, say, 16 to 55 years. Subsequently, he will compare the rate of production obtained by each age group and finally conclude whether or not age is a factor in enhancement of the production. The age is being manipulated by the experimenter by means of selection in order to determine its effects upon the rate of production. Hence, this constitutes and example of Type-S independent variable. A research or investigation which involves the manipulation of the Type-E independent variable is called *experimentation*, no matter whether it is done in a laboratory or in a natural setting. Likewise, a research which involves the manipulation of the Type-S variable is called *correlation research*. Sometimes, a research is done in which there are no independent variables. Such research is termed as *observation*, e.g., when the researcher wants to study the factors underlying the mating behaviour of dogs, he looks at this behaviour from a distance for some time in order to have an idea regarding the factors underlying the mating behaviour. In this example the researcher or the investigator is not manipulating any variable, rather he is simply observing the mating behaviour to know the underlying factors. This illustrates what we call *observation*.

The independent variable (or the stimulus variable as Underwood calls it) may also be classified on the basis of the nature of the variables. Depending upon the nature of variables, the independent variables may be classified into three categories: *task variables, environmental variables and subject variables*.

Task Variables

The task variables refer to those characteristics which are associated with a behavioural task presented to the subject. It includes the physical characteristics of the apparatus as well as many features of the task procedures. There are different types of apparatus used in an experiment. Some are simple and some complex. The simplicity or complexity of the apparatus is likely to produce a change in the behavioural measure. When, for example, the number of wrong paths (blind alleys) in a particular maze is increased, the process of learning or traversing the maze will be a difficult task for the subject. Likewise, when several more or less similar coloured lights are added to a study of complex reaction time, the reaction time is likely to be longer.

Environmental Variables

Environmental variables refer to those characteristics of the environment, which are not the physical parts of the task as such, but tend to produce changes in the behavioural measures. Noise, temperature, levels of illumination and time of the day are examples of environmental variables. Suppose, for example, that the investigator wants to study how reading speed is influenced by the degree of vagueness in handwriting. It is likely that apart from the degree of vagueness, the reading speed may be influenced by the levels of illumination and noise occurring at the time of reading the materials. The intensity of the light and noise constitutes the example of the enviornmental variables because the variations in the intensity of the light and noise also tend to produce changes in the reading speed. Here intensity of light and that of sound are not the physical parts of the task (handwriting) but tend to produce changes in the behavioural measures (relating speed).

Subject Variables

Subject variables refer to those characteristics of the subjects which are likely to produce changes in the behavioural measures. Sex, age, weight, anxiety, intelligence, etc., are the characteristics of the subjects (animals or human), which may be conveniently termed as subject variables. Subject variables can be divided into two types: the *natural subject variable* and the *induced subject variable*. The natural subject variables are those variables which the subjects carry within themselves before the start of the experiment. Age, sex, intelligence, and anxiety are examples of natural subject variables. Induced subject variables, also known as instructional variables, are those subject variables, which are induced by the experimenter's instruction. Suppose the investigator wants to study the effect of two methods (A and B) upon the problem-solving behaviour. He may instruct one group of subjects with one set of instructions and another group of subjects with another set of instructions. The two different sets of instructions are likely to produce changes in problem-solving behaviour. Likewise, the investigator may want to study learning as a function of ego-involvement. He may, for example, give instructions to one group of subjects that a given task is the test of his ability and to another group of subjects that the task is being used as standard practice followed in the experimental situation. These two types of instructions are likely to produce differences in performance of the task.

Sometimes, the experimenter manipulates a variable only for the sake of controlling its unwanted effect upon the experimental results. Such a variable is *not* an independent variable. For example, suppose the experimenter is interested in studying whether or not the reaction time to light is shorter than the reaction time to tone. If he has decided to conduct the individual experiment, he may give 200 random presentations of each stimulus. Now suppose, for the time being, he makes a systematic presentation of the simulus rather than a random presentation. In the first 100 trials he presents tone and in the remaining 100 trails he presents light and subsequently, comes to the conclusion that the reaction time to light is shorter than the reaction time to tone. One may, here, argue that the obtained result is not due to the difference in stimulus but to the practice effect. The subject might have gained some cumulative experiences with 10 presentations of tone, which helped in making quick reaction to light in the remaining 10 presentations. Obviously, then, by making 200 random presentations he would be manipulating the 'sequence of stimulus presentations'. This variable is manipulated to control or eliminate its unwanted effect upon the dependent variable and not for determining its influence upon the dependent variable. Hence, this is not an indepedent variable. Thus, for a variable to be called as independent variable, it must be manipulated for the sole purpose of producing its effect upon the dependent variable, otherwise it will not be an independent variable. In experimental situations, those variables which are controlled by the experimenter because they are not of direct interest but are likely to produce changes in the behavioural measures are known as *control variables* or *relevant variables*. Think of the difference between control variable and controlled variable. Control variables are also known as *extraneous variables*. For example, if the investigator wants to study the relationship between the height and weight of some children, it is very likely that age will tend to affect the relationship because with advancement in age both height and weight of the children will increase. Age here is the example of the relevant variable because it tends to influence the relationship between height and weight of the children. Likewise, if the investigator wants to study the effect of two teaching methods—A and B—upon classroom achievement, intelligence may be one factor which, if uncontrolled, may produce additional differences in the classroom achievement. It is expected that the intelligent children will benefit more from either method of teaching and may improve their classroom achievement. Thus intelligence is here an example of the relevant variable. There are some variables which have no discernible effect upon the dependent variable. These are known as *irrelevant variables*. Relevant variables, or extraneous variables, are of three types, which must be controlled in any experiment.

1. *Subject relevant variables*: Subject relevant variables are those variables which constitute the characteristics of the subject and are controlled by the experimenter because he does not want to study their effect upon the behavioural measure (or the dependent variable). The Type-S independent variables, which are the characteristics of subjects and lend themselves to manipulation through selection, may be conveniently grouped as the subject relevant variables. Thus age, intelligence, race, aptitude, personality classification, etc., are the examples of the subject relevant variables. The students should note carefully that the same variable may act as an independent variable in one experimental situation and as a relevant variable in another experimental situation. A variable which is manipulated and whose effects are studied by the experimenter is called the independent variable but when the experimenter controls the same variable because he does not want to study its effect upon the behavioural measure or the dependent variable, it becomes an example of the relevant variable.

2. *Situational relevant variables*: The situational relevant variables refer to those environmental and task variables whose effects (on the DV) are controlled by the experimenter because they are likely to produce unwanted changes in the DV. The temperature, levels of illumination, complexity of task, etc., are examples of situational relevant variables. All those independent variables whose effects are unwanted and which can be directly manipulated by the experimenter (that is, Type-E independent variables) are included in situational relevant variables.

3. *Sequence relevant variables*: The sequence relevant variables are those variables which arise from the different ordinal positions that the conditions of the experiment occupy in a sequence. For example, when the same subjects are exposed to two or more than than two conditions of the experiment (that appear in a particular sequence), factors like practice, fatigue, adaptation are likely to influence the behavioural measure (or the DV). In the first condition, the subjects' performance will not be affected by practice but in the last condition, it may be.

Similarly, fatigue may not be influencing the performance in the first condition but it may influence the performance in the subsequent conditions. Practice and fatigue are examples of sequence relevant variables which are usually controlled by the counterbalancing design adopted by the experimenter (D' Amato 1970).

From the point of view of the unit of measurement, there are two important ways of classifying variables.

(i) Whether the unit of measurement is categorical (as we find in case of nominal or ordinal scale) or continuous in nature (as we find in case of interval or ratio scale)

(ii) Whether it is qualitative (as we find in nominal and ordinal scales) or quantitative in nature (as we find in interval and ratio scale)

Variables thus classified on the basis of unit of measurement are called qualitative and quantitative variables, and categorical and continuous variables. There is very little difference between qualitative and categorical, and continuous and quantitative variables.

Qualitative Variables And Quantitative Variables

The qualitative variables refer to those variables which consist of categories that cannot be ordered in magnitude. We cannot make such a statement regarding the qualitative variables, e.g., 'category X possesses higher (or lower) magnitude of the variable than category Y.' Thus the qualitative variables comprise the categories which do not have a quantitative relationship among themselves. Sex, race and religion are examples of qualitative variables because they cannot be ordered in magnitude. Since the qualitative variables cannot be ordered in the magnitude, precise and accurate measurements are not possible. As a consequence, they are least preferred in any scientific investigation. The quantitative variables refer to those variables which are composed of categories that can be *ordered* in magnitude. We can, for example, say that category A possesses greater magnitude of the variable than category B. Intelligence, age, levels of illumination, intensity of sound, etc., are examples of quantitative variables. We can say that group A possesses a higher magnitude of intelligence than group B and older people get fatigued sooner than younger and adult ones. Thus the variables can be ordered in terms of magnitude. With the quantitative variables, precise and accurate measurements are possible because they can easily be ordered in terms of increasing or decreasing magnitude. In psychology and education, fortunately, most of the variables belong to the category of quantitative variables. In sociology, qualitative variables are more common.

Continuous Variables and Discrete Variables

Quantitative variables are further divided into two categories, namely, continuous variables and discrete variables. A continuous variable is one which is capable of being measured in any arbitrary degree of fineness or exactness. Of course, the measurement is subject to the limitations of available tools. Age, height, intelligence, reaction time, etc., are some of the examples of a continuous variable. The age of a person can be measured in years, months and days. To go to even a smaller unit of measurement, age can be expressed in terms of hours, minutes and seconds also. Likewise, reaction time can be measured in seconds and milliseconds. If required, it can also be measured in terms of microseconds. Thus, all such variables which can be measured in the smallest degree of fineness are examples of continuous variables. The discrete variables (also known as *categorical variables*) are those variables which are not capable of being measured in any arbitary degree of fineness or exactness because the variables contain a clear gap. For example, the number of members in a family constitutes an example of a discrete variable. The members of a family may be any number like 5, 6, 7, and so on. No amount of refinement in the measuring instrument can produce a value of $5\frac{1}{2}, 7\frac{1}{2}, 6\frac{1}{2}, 7\frac{1}{4}$ or $6\frac{1}{4}$ members. As a general rule, discrete variables are those variables whose values can be determined by *counting*

(D'Amoto 1970). The number of children in a family, the number of females in a particlar state or district, the number of books in a library, and so on, are some of the examples of discrete variables. In psychology and education, most of the quantitative variables belong to the category of continuous variables. It should, however, be noted that independent variables may be qualitative, quantitative, continuous or discrete. The same holds true for dependent variables. However, in most of the psychological and educational researches, independent variables as well as dependent variables are quantitative and continuous.

Moderator Variables and Intervening Variables

The moderator variables are special types of independent variables (also called the secondary independent variables) which are selected by the experimenter because he suspects that these variables may alter or moderate the relationship between the primary (or main) independent variable and the dependent variable. Thus, the moderate variables may be defined as those variables which are manipulated or selected by the experimenter because they are suspected to moderate the relationship of the independent variable with the dependent variable. For example, suppose the investigator wants to study the relative effectiveness of the lecture method and the demonstration method upon classroom achievement. He may take two groups of students of the same class, one to be taught by the lecture method and another to be taught by the demonstration method. But he suspects that the level of intelligence may be a factor which can moderate the relationship between the two methods (the main independent variables) and the classroom achievement (the dependent variable). He may, therefore, manipulate intelligence by dividing both groups into three subgroups having high intelligence, average intelligence and low intelligence. Thus a total of six subgroups would be formed out of which the three subgroups would be taught by the lecture method and the remaining three subgroups would be taught by the demonstration method. Subsequently, the classroom achievement of these six subgroups may be compared with each other. In the above example, the intelligence is manipulated by the experimenter because he suspects that this factor may alter or moderate the relationship between the independent variables and the dependent variable. The intervening variables are the variables which theoretically exist and tend to influence the behavioural measure. Such variables cannot be seen and/or manipulated by the experimenter and their effect can be inferred from the effects of the independent variables as well as the moderator variables upon the dependent variables. Often such variables are not named in the proposed research and hence, little attention is given to them. Suppose the investigator wants to test the following hypothesis: A person who is allowed to reach his goal displays less aggressive acts than the person who is not allowed to reach his goal. In this hypothesis the DV is the aggressive act, the IV is being allowed or not being allowed to reach the goal, and the intervening variable is the frustration. Thus the intervening variable is the variable which is influenced by the independent variable, the relevant variable and the moderator variable and in turn, affects the dependent variable. The intervening variable can be easily identified by carefully examining the hypothesis and asking the question, "What is it about the independent variable that will cause the predicted outcome?" (Tuckman 1978).

Active Variables and Attribute Variables

From the viewpoint of study design that may be experimental or nonexperimental, there are two sets of variables: active variables and attribute variables. A variable which is manipulated by the experimenter is the active variable and the variable which is not manipulated but measured by the experimenter is the attribute variable or organismic variable. Examples of active variables are reward, punishment, methods of teaching, etc. Some of the examples of attribute variables are age, sex, intelligence, race, anxiety, etc. These variables are human characteristics which have already been determined. They cannot be directly altered or manipulated by the experimenter. When the investigator wants to know whether or not ten-year-old girls are superior to ten-year-old boys in numerical ability, it serves as an example of the study of one organismic variable or attribute variable (sex) upon the dependent variable (the numerical ability).

The above distinction between the active variable and the attribute variable is general and is sometimes confusing as well. There are some variables which can be categorized as attribute variables as well as active variables. For example, anxiety is one such variable. Anxiety can be manipulated by giving a set instruction to the subjects. In this case, it becomes an active variable. Anxiety can also be measured with the help of a scale or test. In this case, it constitutes the example of an attribute variable. Thus the variable of anxiety can be studied either as an active variable or as an attribute or organismic variable. Several other examples can be cited in this way.

DIFFERENCE BETWEEN A VARIABLE AND A CONCEPT

Anybody contemplating to do research must know the difference between a concept and a variable. A concept is a mental image or perception and therefore, its meaning varies markedly from person to person. A variable, on the other hand, is measurable with varying degree of accuracy. A concept cannot be measured whereas the variable can be measured by crude or refined unit of measurement. Thus measurability is the main criterion of the difference between a concept and a variable. Since concept is a subjective impression, its understanding differs from person to person and which, if measured somehow, may produce problems in comparing the responses. Therefore, it is suggested that the concept should be converted into variable so that it can be subjected to measurement, though with varying degree of precision from scale-to-scale. Some example of variables are age, gender, attitude and intelligence, which can be measured whereas excellence, satisfaction, domestic violence are examples of concepts.

METHODS OF MEASURING DEPENDENT VARIABLES

In fact, the researcher selects a dependent variable in the study and attempts to provide an operational definition of it. The dependent variable he will select and the operational definition he will provide depends upon the hypothesis and behaviour he is investigating, the subject population and the perspective that he seeks. For measurement of dependent variables, the following approaches are generally undertaken:

(i) *Direct observation of the targeted behaviour*: The dependent variable may involve direct observation and measurement of the interested behaviour. For example, in a study of eating behaviour, Schachter (1968) measured the amount of food eaten by the subjects as a function of different eating cues present. Likewise, Cunningham et al. (1990) conducted one study in which they studied how mood influences our willingness to help others. They presented subjects with cheerful or sad statements, then observed whether subjects would help a confederate.

(ii) *Indirect measures of unseen internal processes*: Sometimes the behaviour to be observed is more an indirect indication of unseen internal processes. The dependent variable measures an observable response that the researcher thinks is correlated with the unseen process. Therefore, changes in the response are made to make inferences about changes in the unseen processes. For example, several researches have been conducted in which the researchers take physiological measurements of heart rate, breathing, blood pressure and perspiration level to draw inferences about the subject's anxiety or stress level.

The use of indirect measures is especially common in studies of young children because they cannot directly tell us about their experiences. Tronick (1989) conducted one study in which he measured the length of time a child stared at a stimulus or noted when he smiled at it, to infer a positive emotional state produced by a given stimulus. In memory research, the number of words correctly or incorrectly recalled or recognized under different conditions is used to infer cognitive processes of retention and forgetting (Craik & Lockhart 1972).

A very common indirect measure is reaction time—the time a subject takes in responding to a stimulus. On the basis of reaction time to stimuli, that differs along a physical or mental dimension, the researcher uses differences in reaction time to infer underlying cognitive or emotional processes involved therein. For example, Sternberg (1969) measured subjects' reaction time to a 'probe word' that either was or was not a part of a previous list for infering how subjects search through the information in their memory.

(iii) *Judgements about a stimulus*: In this approach to the measurement of dependent variable the subjects are asked to make judgements about a stimulus and then, it is observed how the judgements change as a function of the conditions. Another such approach utilizes a forced-choice technique in which the subjects must select from the possible choices provided. In the field of perception, researchers often manipulate characteristics of a visual illusion and require subjects to tell whether or not they still perceive the illusion (Stuart & Day 1991). Sometimes researchers try to obtain more refined judgements. For example, in studying how people perceive the lapse of time, the subjects may be asked to judge the duration of an interval in seconds after certain types of activities have occurred. In case the judgements are not easily verbalized, subjects may be asked to make some symbolic responses. For example, Block (1978) conducted one experiment in which subjects drew a line to indicate the perceived duration of a time interval.

(iv) *Self-reports*: Another approach to the measurement of dependent variable is to obtain self-reports from the subjects. In such reports subjects provide descriptions about their own feelings, ideas or thoughts. Measurement of dependent variable by self-reports is commonly found in rating scales. For example, Schwartz and Clore (1983) manipulated the description that subjects heard about their physical surroundings and then required them to rate statements about their mood level and causes for it. Likewise, such rating scales can also be used to measure subject's attitudes, their perceptions or their confidence in their memory for an event (Brandsford & Franks 1971).

In this way, we find that there are several approaches to the measurement of dependent variables. Depending upon the hypothesis and variables included in the study, any one or more approaches are employed by the researchers.

IMPORTANT CONSIDERATIONS IN SELECTION OF VARIABLES

In any behavioural research, the investigator tries to take a decision regarding the total pool of the variables which are likely to influence the dependent variable. At this juncture, apart from the selection of the main (or independent) variable, he tries to take a decision regarding the variables to be controlled (that is, extraneous variables) as well as variables to be treated as moderator variables. Tuckman (1978) has recommended three important considerations which should be kept in view by the investigator in making the above decision.

1. *Theoretical considerations*: In choosing a particular variable as a moderator variable, the researcher should try to know how it interacts with the independent variable for producing changes in the dependent variable. How frequent is this interaction? He should also know whether or not the variable is related to the theory with which he is working.

2. *Design considerations*: The variables must be selected in view of the scope of design of the experiment. The variable selected as the moderator variable and the extraneous variable must be such that it fulfils the basic requirements of the experimental design. The independent variable must be such that its manipulation can be easily done within the framework of the experimental design.

3. *Practical considerations*: A researcher should also take into account some practical considerations in selecting variables. He should limit the number of variables to be incorporated in the study because it is not possible to study too many variables at a time. In selecting variables as moderator variables, independent variables and extraneous variables, he should also keep in view the financial resources available for the purpose. The time consideration should also not be ignored. He should also see the nature of the variables because some variables are easy to study and some are difficult. Above all, in selecting the variables he should also be aware of the control he is likely to have over experimental situations. When the variables are difficult to study, he is likely to lose control over the experimental situation.

In selecting variables for any investigation the above considerations, if kept in view, can be useful and are likely to increase the validity of the research.

IMPORTANT APPROACHES TO MANIPULATING INDEPENDENT VARIABLES

Manipulation of independent variables in experimental researches is very important. There are several ways of manipulating independent variables. However, the important approaches or ways of such manipulation are given below.

(i) *By manipulating context*: Often the researcher manipulates the context in which the stimulus is presented in order for manipulating the independent variables. Researchers studying learning processes have generally examined the effect of different types and amounts of rewards or punishments on the subjects' ability to learn. Some other researchers have induced different types of moods through hypnosis and have then examined its effect upon subjects' memory and thought processes. In all these examples, the independent variables are shown to be manipulated through manipulating the context.

(ii) *By presenting different stimuli*: Researchers often manipulate the independent variables by creating various conditions of an independent variable by presenting subjects with different stimuli or changing the characteristics of a stimulus. For example, in studying aggression, Green (1978) exposed subjects to films in an experiment in which different amounts of violence were depicted in different conditons and its impact upon the aggressiveness subsequently shown by the subjects was examined. Likewise, researches in the field of applied social psychology have measured the effect of various furniture arrangements on social interactions (Campbell & Harren 1978).

(iii) *By manipulating the likely information to be given to the subjects*: Sometimes the independent variable may consist of the instructions or information given to the subjects in each conditon. These instructions are itself varied and their impacts are examined. For example, Loftus (1975), for studying memory, examined how different information conveyed in questions about an event affected the subject's recall of that event. Likewise, Phesterson, Kiesler & Goldberg (1971) conducted one study to examine whether the subjects would evaluate the painting in a sex-biased manner in which the subjects were presented a painting and were told that the supposed painter was a male or female.

(iv) *By manipulating social setting in which confederates are used*: Some researchers have tried to manipulate independent variables by manipulating the social settings in which help is taken from confederates who are people entrusted by a researcher to act as other subjects or accidental passers-by so that a particular type of social situation may be created and the real subject may respond to this. For example, Asch (1951) conducted one experiment in which several confederates were present and a subject, asked to judge the length of a line, drastically overestimated its length in order to show conformity to the social pressure being exerted by a group of confederates.

(v) *By manipulating intervening variables*: Intervening variables are internal states that influences a behaviour. In fact, it intervenes or comes between independent and dependent variables. The intervening is enhanced by the independent variable, which in turn, influences the dependent variable. For example, if we frustrate the subjects by preventing them from obtaining a desirable reward, it will lead to a state of anger which, in turn, leads to greater amount of aggressive behaviour. Here, the independent variable like frustration is being manipulated by preventing the subjects from obtaining a desirable reward. In doing so, we presumably change the intervening variable of anger which is likely to influence the subject's scores on the dependent variable of aggressiveness.

Here it would be profitable to make distinction between two types of internal psychological states—*a state characteristic* and *a trait characteristic*. A state characteristic is a temporary, changeable attribute that is influenced by situational factors. For example, state anxiety is a state characteristic and the level of anxiety the person will experience depends upon the situation he is in. Conversely, a trait characteristic is something stable over time and easily influenced by situational factors. For example, self-esteem is generally defined as a trait characteristic because it is not influenced by the situation in which the person is in. It is important to remember that state characteristics can be manipulated and studied as intervening variables because they can change in response to the conditions of independent variable. Trait characteristics cannot be studied in this way because they are not influenced by most manipulations due to its nature of permanence.

(vi) *By stressing or overloading psychological system*: In some researches, the researchers try to manipulate the independent variable by overloading or stressing a psychological system and then, inferring from the subject's responses how the system normally works or operates. For example, in one study subjects were socially or emotionally stressed by having them produce a behaviour that conflicts with an attitude they advocate and measuring how they resolve such conflict (Sherman & Gorkin 1980). Likewise, in one study of attention the researchers presented two 'dichotic messages' simultaneously, one to each ear. In this case, the independent variable consisted of varied instructions or message content and the dependent variable was the way in which each message was heard (Gray & Wedderburn 1960).

(vii) *By manipulating physiological processes*: In some researches, the researchers manipulate the independent variable by manipulating the physiological process of the organism. For example, in many animal researches the researchers may employ different surgical techniques to create different conditions in which some parts of the brain are removed or cut (Tokunaga et al. 1986). The dependent variables in such studies were the impact of such removal or cut upon the animal's behaviour, motivation, memory, and so on.

In this way, we find that independent variables in psychological researches are manipulated in different ways. Depending upon the nature of variables, the researchers choose any one or two methods of manipulation for their research work.

TECHNIQUES OF CONTROLLING EXTRANEOUS VARIABLES

As we know, the extraneous variables are those that operate in the experimental situation in addition to the independent variables and affect the dependent variables. It is, therefore, essential that extraneous variables must be controlled. If the researcher fails to control the extraneous variables, it results in a confounded experiment. The following are the five important ways to control the extraneous variables:

(1) Technique of elimination
(2) Constancy of conditions
(3) Balancing
(4) Counterbalancing
(5) Randomization

These methods can be discussed as follows:

Technique of Elimination

The simplest way to control the extraneous variable is to eliminate it completely from the experimental situation. For example, if noise in any experiment is an extraneous variable, the simple way to control it is to make the experimental situation soundproof.

But this technique is too simple to control many extraneous variables. For example, controlling extraneous variables like age, sex, intelligence through the technique of elimination is very difficult.

Constancy of Conditions

Where the extraneous variables can't be controlled through the technique of elimination, they can be controlled by holding their values constant for all participants in all conditions. This is known as constancy of conditions. By holding constant the instructions to be given to every participant, by holding the time of the day constant for all participants, and by holding lighting conditions constant, extraneous variables related to instruction, time of the day and lighting conditions can be controlled. The apparatus for administering the experimental treatment and for recording the results should also be kept constant in all conditions for all participants. Sometimes the organismic variables like sex, age, intelligence also become important extraneous variables. To control these extraneous variables, the researcher chooses only those participants or subjects who are homogeneous with respect to either sex or age or intelligence. Using all participants of the same sex naturally controls the effect of sex, if any, upon dependent variables.

Balancing

When due to some reasons, it is not possible for the researcher to hold the various conditions constant, he may try to control the extraneous variables by balancing a technique used in the following two situations:

(i) where the researcher remains unable to identify the extraneous variables
(ii) where they are readily identified and the researcher takes special steps to control them

Let us consider the first situation where the researcher has no idea about the likely extraneous variables that might be influencing the dependent variable. In such a situation the researcher gives equal treatment to all subjects, but the experimental group is treated in a different way from the control group. Consequently, wherever the extraneous variables operate, they influence both the experimental group and the control group in equal manner and their effect is, thus, balanced. The changes occurring in the dependent variable are clearly attributed to the changes done in the independent variable. Suppose a set of four extraneous variables is influencing the experimental group in addition to the independent variable. The effects of these extraneous variables can be balanced out by allowing them to operate also on the control group. As a consequene, the independent variable will be the only one that can differentially influence the two groups (cf. Figure 20.1).

Another situation of balancing is that in which the experimenter knows well about the extraneous variables that are of influence. Suppose the researcher knows that sex is the extraneous variable in a particular experiment. To control the effect of sex by balancing, the experimenter may assign an equal number of subjects from each sex to each group. This will balance out the effect of sex upon dependent variables. Likewise, if age is an extraneous variable in an experiment, he may assign an equal number of each age classification to each group.

Thus balancing helps in controlling the extraneous variable. The obvious assumption of balancing is that the participants in the control group and the experimental group are initially equivalent. They are only treated differently.

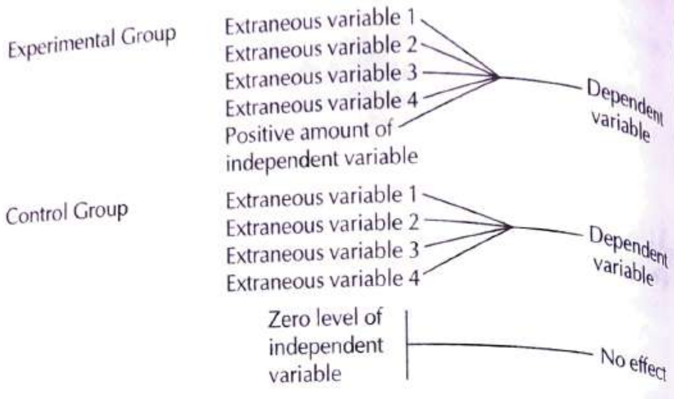

Fig. 20.1 Illustration of the use of the control group as a technique of balancing

Counterbalancing

The technique of counterbalancing is used for controlling the extraneous variables of amount of practice and fatigue. There are such experiments in which each subject is required to serve under two or more different experimental conditions and there is the probability that the participants' performance might improve due to practice or might decrease due to fatigue. The technique of counterbalancing is used to distribute these practice and fatigue effects together called *order effects*, equally over all conditions. Therefore, whatever their effects, they tend to influence behaviour under each condition equally since each condition occurs equal number of times at each stage of practice. Thus the general principle of counterbalancing is that each condition must be presented to each subject an equal number of times and each condition must occur an equal number of times at each practice session. Besides, each condition must precede and follow all other conditions an equal number of times.

When the experimenter tries to control the order effects by means of counterbalancing, he assumes that there is occurring a *differential transfer* or *asymmetrical transfer* which needs to be controlled. Asymmetrical transfer is the transfer in which transfer from condition A (when it occurs first) to condition B is different from the transfer from condition B (when it occurs first) to condition A. Table 20.1 illustrates the symmetrical transfer and asymmetrical transfer. In Table 20.1 we find that there are two conditions of study, A and B. In a symmetrical transfer those who received condition A first obtained a score of 6 on dependent variables and when subjected to second experimental condition B, they got a score of 9. The difference was three.

Table 20.1 Symmetrical and Asymmetrical transfer between conditions A and B

	Symmetrical transfer			Asymmetrical transfer		
	A	B	Difference	A	B	Difference
One half of subjects	6	9	3	6	9	3
	B	A		B	A	
One half of subjects	4	7	3	3	8	5

The other half of the subjects received a score of 4 in condition A and a score of 7 in condition B. The difference was again 3. Thus difference for both the groups was the same and hence, the transfer was symmetrical, that is, no differential transfer was found. But in the asymmetrical transfer, the difference for one half of subjects from going to A and B was 3 but for the other half of subjects from going to A and B was 5. In symmetrical condition it would be legitimate to combine the values of A (6 + 7) and compare it with the values of B (9 + 4), but in asymmetrical condition it will not be possible for the researcher to combine the values of A and compare it with the values of B. Counterbalancing is done for controlling asymmetrical transfer. If the design is perfectly counterbalanced, the transfer will be a symmetrical one.

The difference between *balancing* and *counterbalancing* is sometimes confused. In fact, the two are different. The experimenter preferably resorts to counterbalancing when each subject receives more than one treatment (AB or BA) and an attempt is made to distribute the fatigue and practice effects equally all over the experimental conditions. In *balancing*, this is not the case because here each subject receives only one experimental treatment and the effects of extraneous variables are balanced out by affecting the members of both the experimental group and the control group equally.

Randomization

Randomization is a very popular technique of controlling extraneous variables. Randomization refers to a technique in which each member of the population or universe has an equal and independent chance of being selected. Randomization is used where the experimenter assumes that some extraneous variables operate, but he can't specify them and, therefore, can't apply the other techniques of controlling extraneous variables. This technique is also applied where the extraneous variables are known but their effects can't be controlled by known techniques.

The importance of randomization lies in the fact that this technique randomly distributes the extraneous effects over the experimental and control conditions. Such balancing occurs whether or not the experimenter has identified certain extraneous variables, because the effects of unknown or unspecified extraneous variables are said to be equally distributed across different conditions of the experiment when the experimenter randomly assigns subjects to the different groups or conditions.

Suppose the experimenter wants to study the effect of lighting upon the reading ability of the subjects and for this, he tests two groups under two different conditions. The experimental group is tested in the morning conditions where subdued lighting is found and the control group is tested in afternoon conditions where bright light is found. The experimental group had been given training on how to read correctly. In such a situation it is expected that the reading ability score for the experimental group will be inferior to the reading ability score of the control group (because of subdued lighting) despite the training. But if each subject has an equal opportunity to serve in the morning or afternoon then on the average, the lighting will affect both the groups equally and confounding due to lighting conditions will be automatically eliminated.

In this way, it is obvious that the extraneous variables can be controlled in several ways. Of these various ways, randomization, balancing and counterbalancing are relatively more popular.

CONTROLLING DEMAND CHARACTERISTICS

In behavioural study, sometimes the social and physical surroundings in the situation provides some cues that essentially demand that we should behave in a certain way and not behave in some other way. In research, these cues are called *demand characteristics*. In other words, demand characteristics are defined as cues provided by research context that guides or biases a subject's behaviour in a particular direction. In fact, subjects try to depend upon these demand characteristics for answering questions like what are they supposed to do? What is really going on here? What meanings will be derived from my response?, etc.

Demand characteristics can be explained in terms of the following four components of a study: *subjects, environments, measurement procedures* and the *experimenter.*

Subjects bring with them certain kinds of attitudes and biases that influence their behaviours. While subjects try to do their best as instructed in the study, they are also on vigil and remain sensitive to any such display which may term their behaviour abnormal. Because of these concerns and apprehension, different subjects react differently to the same situation. Some subjects behave very cooperatively and do what they are supposed to do whereas some other subjects may behave in a defensive way. They may not feel good but intentionally act being happy.

Environment in which the study is carried out, also feeds such reactions among subjects. Such environment may distract subjects from the main task or cause them to react in a certain biased way. In environment, random noises, changes in the level of illumination or broken pencil points are interpreted differently by different subjects and leads them erroneously to conclude about what is being studied and how they should respond in such situations.

Measurement procedure also produces demand characteristics in light of which subjects tend to modify their experiences during study. For example, in studies of imaging, subjects are required to envision a previously presented map and the time that they take in mentally traveling to different locations are measured. Generally, subjects take longer a time when locations on the map are distant than when they are nearer. This result may occur not because it reflects how subjects travel along mental maps but because subjects know fully well that longer distances should take longer and therefore, they 'oblige' the experimenter.

The *experimenter* is also one important source of demand characteristics during study. The experimenter generally dresses and behaves formally to seek maximum cooperation from the subjects. But such formality may inhibit many of the subject's normal reactions. Not only this, if the experimenter acts and dresses informally, it may encourage inattentiveness and poor performance. Apart from this, subjects are very sensitive to the experimenter's expectancies and therefore, any subtle cues about the response provided by the experimenter affects the behaviour of the subjects. Such expectancies occur because the experimenter, as researcher, knows about the prediction of the study and may inadvertently communicate them to the subjects, thus affecting their behaviour.

In this way we find that there are several components of demand characteristics that, in fact, modify the subject's behaviour. What are the *impacts* of these demand characteristics? The obvious impact is that the subjects respond to these cues or characteristics rather than to respond solely to the treatment given in the condition. As a consequence, we lose both the internal and the external validity of the study. Internal validity is lost because the subjects behave in the predicted manner, making it appear that our manipulation influences their behaviour, whereas the reality is that the cues provided by the demand characteristics themselves make subjects behave in that way. Sometimes it may also happen that cues may cause subjects to behave contrary to the predictions, making it appear that the manipulation does not influence behaviour. External validity is also lost because when the subjects don't respond to manipulation and therefore are not behaving naturally, the obtained results don't generalize to other situations in which the demand characteristics are absent. Apart from validity, demand characteristics also reduce reliability because not all subjects react to the cues in the same way and then in this case, inconsistency occurs causing greater error variance in the scores of the subjects.

Now the question arises: How can we *control* these demand characteristics? To answer this question, the following measures are usually taken:

(i) *General strategies of control for demand characteristics:* Researchers have evolved *three* general strategies for controlling demand characteristics. The first strategy is that the researcher should provide as few cues as possible. If the subjects have no cues, or fewer cues they will behave in a natural manner. For providing little cues the experimenter should not divulge the specific purpose, should not include extraneous, distracting information and should avoid extraneous behaviours or comments and should hide the threatening equipment and apparatuses.

Second, the general strategy is to neutralize the present cues to the extent it is possible. The researcher tries to be neither friendly nor unfriendly and keeps his position neutral. Thus the experimenter presents the task without implying that it is difficult or easy. He also does not indicate what is a normal response. He also tries to neutralize the subject's fear and suspicions.

The third general strategy is to seek experimental realism which refers to the extent to which the experimental task engages subjects psychologically such that they become less concerned with demand characteristics. For establishing experimental realism, the experimenter defines the independent variable and other related aspects of the experimental task in such a way that the subjects may find the task interesting and challenging. In this way, they will 'forget' about the demand characteristics and will be realistically engaged in the task.

(ii) *Use of deception for controlling demand characteristics:* Deception is another way of controlling demand characteristics. Deception refers to the creation of an artifical situation or a 'cover story' for disguising the procedures of the study. In such a situation subjects are not aware of the independent variable being manipulated and then, they don't feel pressurized to respond in a certain way. Deception is used when a straightforward presentation of a measurement procedure produces such strong demand characteristics that cannot be eliminated or neutralized in a proper way. Sometimes, the experimenters use demand characteristics to their vantage, disguising a task by incorporating a procedure that may look scientific. For example, Duclos et al. (1989) conducted an experiment in which they manipulated the posture adapted by the subjects in order to determine how posture influences their mood. In this study, 'fake electrodes' for measuring brain activity were attached to them and they were told that their posture was important for accurate and precise measurement.

(iii) *Placebo to control demand characteristics:* Placebo is also used to control the demand characteristics. If the design of the study is such that it uses both an experimental group as well as a control group, placebo is used to control the demand characteristics. What happens that if the experimental group is exposed to stimuli, tasks or instructions which are not presented to the control group then a confounding occurs: here the researcher is not sure whether the experimental group behaves differently because of the treatment or because of the demand characteristics that accompany or follow it. It can be explained through an example like this: Suppose the experimental group is given alcohol and the control group is not. Any impairment in behaviour that the experimental group demonstrates may be due to alcohol or it may be due to the fact that giving subjects alcohol implies that they will behave drunkenly and so they do. To control such demand characteristics, the control group is given a placebo, that is, this group will also be given something that will be called alcohol whose smell and taste will be that of alcohol but in reality it will contain zero amount of alcohol. In such a situation, any difference in behaviour between the control group and the experimental group can be attributed to the real alcohol given to the experimental group. Thus we see that placebos are given to control groups so that they will experience the same demand characteristics as the experimental groups and that way, the effect of demand characteristics is equalized for both groups and become controlled.

(iv) *Controlling expectancies of subject and experimenter:* It is a well-known fact that the expectancies of the subject and the experimenter influence the outcome of the study by way of introducing various kinds of demand characteristics. For controlling the subjects' expectations, *single-blind procedure* is adopted in which subjects are unaware of the specific condition they are receiving. Likewise, for controlling experimenter's expectations, *double-blind procedure* is used in which both the researcher and the subjects are unaware of the specific condition being presented. The original experimenter trains other researchers to actually conduct the study but they don't know and that way remain blind to the condition a subject is in and to the various predictions of the study. Suppose for example, the experimenter is interested in knowing the influence of a drug or medicine. If he knows to which group of subjects a placebo or actual drug is being administered, he may directly or indirectly communicate his expectations to the effect of such treatment to the subjects. These expectations alone can produce various kinds of physical reactions, etc. However, if the experimenter does not know or is blind to such information, no expectations will be communicated to the subjects that may influence the study's outcome.

Thus we find that there are several ways of controlling demand characteristics which, if remain uncontrolled, do influence the outcome of the study seriously.

☑ Review Questions

1. Discuss the various kinds of variables and indicate their importance in psychological research.
2. State the importance of control in an experimental study of behaviour. Describe with illustrations the different techniques that have been devised for controlling extraneous variables.
3. Differentiate between constructs and variables. Giving suitable examples, elucidate the different types of variables.
4. Discuss the important considerations in selection of variables. Illustrate through examples the importance of balancing and counterbalancing in the control of extraneous variables.
5. Discuss, with examples, the various ways of manipulating independent variables.
6. Discuss the various methods of measuring dependent variables.
7. Describe the various techniques of controlling demand characteristics in an experiment.
8. Write short notes on the following:
   (a) Type-E and Type-S independent variable
   (b) Demand characteristics
   (c) Extraneous variable

◆

21

RESEARCH DESIGN

CHAPTER PREVIEW

- Meaning and Purpose of Research Design
    - Experimental Variance
    - Extraneous Variance or Control Variance
    - Error Variance
- Criteria of Research Design
    - Capability to Answer Research Questions Adequately
    - Control of Variables
    - Generalizability
- Basic Principles of Experimental Design
    - Replication
    - Randomization
    - Locus Control
- Basic Terms Used in Experimental Design
- Some Important Types of Research Design
- Between-Subjects Design
    - Two-Randomized-Groups Design
    - More-Than-Two-Randomized-Groups Design
    - Matched-Groups Design
    - Which Design to Prefer: Randomized-Groups or Matched-Groups
    - Factorial Design
- Problem of Creating Equivalent Groups in Between-Subjects Design
- Within-Subjects design
- Problem of Controlling Sequence Effects in Between-Subjects Design
- Comparison of Between-Subjects Design and Within-Subjects Design
- Experimental Design Based upon the Campbell and Stanley Classification
- Pre-Experimental Design (Nondesigns)
    - One-Shot Case Study
    - One-Group Pretest–Posttest Design
    - Static-Group Comparison (or Intact-Group Comparison)
- True Experimental Designs
    - Posttest only, Equivalent Group Design
    - Pretest–Posttest Control Group Design
    - Solomon Four-Group Design
    - Latin Square Design
- Quasi-Experimental Designs
    - Time-Series Designs
    - Equivalent Time-Samples Design
    - Nonequivalent Control Group Design
    - Counterbalanced Design
    - Separate-Sample Pretest–Posttest Design

  - Patched-up Design
  - Longitudinal Design
  - Cross-Sectional Design
  - Cohort Design
- Ex Post Facto Design
  - Correlational Design
  - Criterion-Group Design
- Steps in Experimentation

MEANING AND PURPOSE OF RESEARCH DESIGN

Research design is the detailed plan of the investigation. In fact, it is the *blueprint* of the detailed procedures of testing the hypotheses and analyzing the obtained data. Research design, thus, may be defined as the sequence of those steps taken ahead of time to ensure that the relevant data will be collected in a way that permits objective analysis of the different hypotheses formulated with respect to the research problems. Thus, research design helps the researcher in testing the hypotheses by reaching valid and objective conclusions regarding the relationship between independent and dependent variables. The selection of any research design is obviously not based upon the whims of the researcher, rather it is based upon the purpose of the investigation, and types of variables and the conditions in which the research is to be conducted.

The purpose of any research design is to provide a maximum amount of information relevant to the problem under investigation at a minimum cost. Basically, a research design serves two functions.

First, it answers the research questions as objectively, validly and economically as possible. In fact, this is an important function served by any research design. The research problems are usually epitomized by the hypotheses. A research design suggests to the researcher how to collect data for testing these hypotheses, which variables should be treated as control variables, what methods of manipulation will be more adequate in a particular context, what types of statistical analyses should be done and finally a possible answer to the research problems. Thus a research design, after moving through the sequence of different related steps, enables the researcher to draw a valid and objective answer to research problems.

Second, a research design also acts as a control mechanism. In other words, it enables the researcher to control unwanted variances. In any scientific investigation there are three types of common variances, namely, the *experimental variance*, the *extraneous variance* and the *error variance* with which the researcher is directly concerned. A discussion of these variances is presented below.

Experimental Variance

Experimental variance is produced in the dependent variable by the manipulation of the experimental or the independent variables by the experimenter. Usually, the experimenter or the researcher wants to maximize the experimental variances so that he can get as valid and objectve data as is possible. For maximizing the experimental variance, the investigator designs the experiment in a way in which the different experimental conditions become as different as possible. Suppose the experimenter is interested in studying how the rate of learning a task is influenced by the differential reward. This simple experiment may be conducted by randomly placing the subjects matched on age, sex, intelligence, etc., (all these here constitute examples of the relevant variables) into three experimental conditions: high reward, medium reward and low reward. After the subjects in each experimental condition have learnt the task, the learning rate may be determined in terms of, say, the trials taken to reach the criterion. Fewer trials taken in reaching the criterion may indicate higher speed of learning the task. If it comes out

that the condition for the high reward is producing a higher rate of learning and the condition for the low reward is producing a poor rate of learning, the experimental variance is at maximum. Suppose for the time being that the experimenter designs the experiment in a way in which all the three conditions are nearly equal (say, for example, all three conditions are either high-or-low reward conditions), the experimental variance will be at minimum because no discernible changes are expected in the learning rate by the three conditions. Thus the higher the difference between the experimental conditions, the higher the experimental variance. The experimenter always tries to maximize the experimental variance.

Extraneous Variance or Control Variance

Extraneous variance or control variance is produced by the extraneous variables or the relevant variables. An experimenter always tries to control the relevant variables and thus, also wants to eliminate the variances produced by these variables. For elimination of extraneous variance, it is essential that the extraneous variables be properly controlled. There are five such ways which have been discussed in Chapter 20.

Error Variance

The third function of a research design is to minimize the error variance. Error variance is defined as those variances or variabilities in the measures which occur as a function of the factors not controllable by the experimenter. Such factors may be related to the individual differences among the subjects themselves such as to their attitude, motivation, need, ability etc., or they may be related to what is commonly called the errors of measurements such as the differences in trials, differences in conditions of experiment, temporary emotional state of the subject, fatiguability, etc. Whatever the source of error variance, it has three distinct features.

1. Error variance is self-compensating because sometimes the variability is positive and sometimes negative.

2. Since error variance is both positive and negative, it tends to cancel out in several repeated measurements. Hence, it can be said that in several repeated measurements the mean of the error variance will be zero.

3. Error variance is unpredictable probably because it is based on random errors. In this way, error variance is distinct from systematic variance because the latter is predictable and based upon systematic errors.

The minimization of error variances serves basically two purposes. *First*, it improves the reliability of measures so that the generalization can be more dependable and accurate. Thus it strengthens the external validity of the experiment. *Second*, it gives the systematic variance a chance to show its significance, if it is really significant. If the error variance (in ANOVA, usually the within-group variance) is large, the systematic variance (or between-variance) will not be significant.

The error variance can be usually minimized by controlling the conditions of the experiment. If the different conditions of an experiment are fully controlled, the occurrence of error variance is minimized to a great extent. In an experiment allowing many uncontrolled conditions, the error variance accumulates to higher proportions.

Finally, it may be added that one of the objectives of a good research design is to provide unambiguous results and to avoid 'confounding'. In the technical sense, the term 'confounding' refers to that state in an experimental situation where two or more variables have changed concurrently so that the resulting change in dependent variable measures cannot be attributed to a single variable. Obviously, confounding of variables introduces vagueness in stating a cause–effect (or independent–dependent variable) relationship and hence, a good research design also purports to avoid such confounding of variables.

CRITERIA OF RESEARCH DESIGN

As we know, research design is data discipline. There are various types of research designs. Some are weak designs while some are strong. Behavioural researchers have been able to formulate certain criteria on the basis of which a weak research design can readily be distinguished from a strong one. These criteria have proved very useful in guiding the researchers in the correct direction. These criteria are mentioned below.

1. Capability to answer research questions adequately
2. Control of variables
3. Generalizability

A discussion of these research designs follows.

Capability to Answer Research Questions Adequately

A good research design is the design that answers research questions adequately. Sometimes the researcher selects a design which is not appropriate for answering the research question in hand. Such designs constitute an example of weak research design. Such a design does not adequately test the hypothesis either. It is a common practice that students, while trying to answer a research question by conducting experiment or doing research, often match sex, age and intelligence of the subjects on the assumption that such matching would lead to the setting of a better experimental group and control group. The reality is that if there is no relation between say, age and the dependent variable then matching an age will be irrelevant. Therefore, any design based upon matching would be a weak design.

Likewise, if the research problem is such where setting of four groups, namely, three experimental groups and one control group or four experimental groups are necessary but an researcher takes only two groups and utilizes a two-group randomized design, this will again be an example of a design which will not answer the research questions adequately. Likewise, for testing interaction hypothesis a factorial design would be an appropriate one, whereas a two-group randomized design would be an example of an inappropriate design.

Control of Variables

Another criterion of a good research design is that it should control the effects of extraneous variables which are more or less similar to independent variables that have the capacity to influence dependent variables. If left uncontrolled, such variables are called *independent extraneous* variables or simply *extraneous variables*. A design which fails to control the effect of extraneous variables is considered a weak one and the researcher should avoid such a design.

There are various ways to control the effects of extraneous variables. Of these ways, randomization is considered by many as one of the best techniques of controlling the extraneous variables. There are three basic phases in randomization—random selection of subjects, random assignment of subjects into control and experimental groups and random assignment of experimental treatments among different groups. Sometimes, it happens that it is not possible for the researcher to make a random selection of subjects. In such situations, the researcher tries to randomly assign the selected subjects into different experimental groups. When this random assignment is not possible due to any reason, the researcher randomly assigns the different experimental treatments among experimental groups. Whatever the method may be, randomization has proved very useful in controlling the extraneous variables. A design which fully controls the extraneous variable is considered to be the best design for the research. This increases the internal validy of the research.

Generalizability

The third criterion of a research design is generalizability. Generalizability is the external validity of the research. In other words, it refers to the extent to which the results of the experiment or research obtained can be generalized to subjects, groups or conditions not included in a sample of the research. If the design is such that the obtained results can be generalized to larger groups or subjects, the design is considered to be a good one. In fact, how much the researcher can generalize the results obtained is a complex question and it is concerned not only with technical matters of the research like sampling and research design but is also concerned with the larger problems of basic and applied research. In basic or fundamental research, the problem of generalizability is not serious because this is not the first consideration. Here the researcher is basically concerned with examining the relation among variables and exploring why the variables are so related. But in applied research, the researcher's main concern is with generalizability because the researcher intends to apply the obtained results to other persons, conditions or situations.

These are the basic criteria of distinguishing a good research design from a weak research design.

BASIC PRINCIPLES OF EXPERIMENTAL DESIGN

Having discussed the different types of research design, we are now in a better position to concentrate upon the basic principles of *experimental investigation*. The base of all experimental investigation is the experimental design, which is one of the types of research design. An experimental design may be defined as simply a sequence of steps (taken ahead of time), which permit the objective analysis of objective data in a way that a definite cause-effect relationship can be inferred between the independent variable and the dependent variable. As earlier discussed, an independent variable is defined as the variable, which is manipulated by the experimenter either directly or through selection, so that its effects can be studied upon the behavioural measure. The variable is so named because we can change it *independently*. A dependent variable is the response variable or the behavioural measure, which is affected by the manipulation of the independent variable. The variable is so named because responses depend upon those changes or manipulations. Suppose, for example, the experimenter is interested in studying the effect of the intensity of illumination upon retinal fatigue. Obviously, in this situation the experimenter would change the levels of intensity of illumination and would see how long the subject is able to read under each level. In this example, the intensity of illumination is the independent variable and retinal fatigue is the dependent variable.

A sound experimental design is based upon some principles. According to Ostle & Mensing (1975, 260), there are three basic principles of an experimental design: *replication*, *randomization* and *local control*.

Replication

The term 'replication' is really a fusion of two words, namely, *duplication* and *repetition*. It refers to the deliberate repetition of an experiment, using a nearly identical procedure with a different set of subjects, in a different setting and at a different time. Winer (1971, 391) has similarly said, "A replication of an experiment is an independent repetition under as nearly identical conditions as the nature of the experimental material will permit." Replication permits a person in revalidating a previous study or raises some questions about the previous studies. Thus, replication provides a very accurate estimate of the *experimental error* that can be used as a basic unit of measurement in evaluating the significance of the observed differences. The term 'experimental error' is very frequently used in the literature of experimental design. Hence, it needs a more detailed explanation. Experimental error refers to the errors occurring due to faulty experimental design, faulty measurement, biased observation, uncontrolled variations among the experimental units (or subjects) and the uncontrolled extraneous variables. At the end, it may also be added that replication should not be *confused* with multiple measurements. Suppose the experimenter selects a particular level of intensity of light and experiments its effect upon the retinal fatigue of several subjects one after another. This is obviously not a case of replication but a case of multiple measurements.

Randomization

Randomization is the second basic principle of experimental design. It makes the test valid. How? As we know, each statistical test used in an experimental situation depends upon some assumptions of which the most frequent and common assumption is that the observations should be independent. The independence of the observations is maintained when the samples are randomly drawn from the population or the subjects are randomly assigned to the experimental treatments. Thus randomization ensures the independence of the observations which, in turn, makes a statistical test valid. Not only this, when subjects are randomly assigned to the experimental treatments or the experimental treatments are randomly assigned to subjects, this automatically controls the extraneous variables, which, otherwise, are left uncontrolled. Thus, randomization is like an 'insurance' which is "always a good idea, and sometimes even better than we expect" (Ostle & Mensing 1975). Sometimes it has been found that complete randomization becomes difficult. This is especially true when one is dealing with organismic variables. In such a situation, the experimenter should not insist on complete randomization but a middle path between complete randomization and complete nonrandomization or systematization should be followed. A very encouraging suggestion has been made by Ostle & Mensing (1975, 264): "There are situations in which complete randomization is either impossible or uneconomical.The statistician should not, therefore, adopt the unrelenting position of insisting on complete randomization in every case. On the other hand, neither should he agree to the use of a completely systematic design ... Some intermediate position ...is often most realistic."

Local Control

By local control we mean the amount of balancing, blocking and grouping of the subjects or the experimental units employed in an experimental design. Thus local control is, in a way, regarded as a part of the entire edifice of the experimental design. These three terms, namely, 'balancing', 'blocking' and 'grouping' need further explanation. The term *grouping* is most easy to define. It refers to the assignment of homogeneous subjects or experimental units into a group so that different groups of homogeneous subjects may be available for differential experimental treatments. *Blocking* refers to the assignment of experimental units to different blocks in such a manner that the assigned experimental units within a block may be homogeneous. *Balancing* in an experimental design refers to the fact that grouping, blocking and assignment of experimental units to the different treatments have been done in such a way that the resulting design appears to be a balanced one. A design to be statistically and experimentally sound must possess the property of local control.

BASIC TERMS USED IN EXPERIMENTAL DESIGN

Experimental design uses a defined set of terminology. It is essential to get acquainted with these terminologies. These terminologies are discussed below.

(a) *Factor:* A factor is a variable which is defined by the experimenter. A factor is also used interchangeably with the terms *treatment* or *experimental variable*. A factor is also referred to as *independent variable*, which is manipulated by the experimenter. This is called as experimental variable. For example, suppose the experimenter manipulates the intensity of sound to examine its impact upon performance of a sensory-motor task. Here sound is an experimental variable and is referred to as *treatment factor*. Such type of factor is called *E-type of factor* manipulated directly by the experimenter. Besides this, there may be S-type of factor, which is manipulated by selection. In fact, these are *subject-related variables*, which can't be directly manipulated by the experimenter. For example, if the experimenter is interested in studying the impact of age on visual acquity, he may manipulate the age by selecting subjects of different age levels. When a variable is manipulated by selection, it is generally called as a *classification factor*. In factor analysis that tells us which variables tend to clump together, that is, which ones tend to be correlated with each other and not with others, each clump (group of variables) is called a *factor*.

Factors in an experimental design are usually denoted with capital letters such as *A*, *B*, *C*, *D*, etc. For example, in the experiment trying to examine the impact of levels of illuminance and noise upon performance of letter cancellation task, level of illuminance may be designated as *A* factor and level of noise as *B* factor.

(b) *Levels:* By level is meant each specific variation in a factor. For example, the factor sound may consist of three levels: 40 d_B, 60 d_B and 80 d_B. The experimenter tends to choose the number of levels of a factor in either of the two ways: by adopting some systematic, nonrandom procedure and by random procedure. Some factors may have infinite number of potential levels (such as intensity of sound) and others may have only a few (from example SES, sex of the subject). In case the experimenter decides to select *p* levels from potential *P* levels by using some systematic, nonrandom procedure, this factor is considered to be a fixed factor. This is known as *fixed effect model* where the chosen levels are the ones in which we are interested and the experimenter does not wish to generalize beyond these levels of the factors. Thus if one wants to replicate the experiment, the same set of treatments/levels will be included in each replication and the drawn conclusions are restricted to those particular experimental treatments or levels.

In contrast to the systematic and nonrandom procedure, the experimenter may decide to include the *p* levels from potential *P* levels through random procedure. In such a situation, this is considered as *random factor* and this is called as *random effect model* or *variance component model*, which is rarely used. This model requires that the population of levels should be defined and then, selection of several levels be made randomly. In behavioural sciences, this model seems unrealistic because here the experimenters deliberately choose certain levels of factors for their experiment.

(c) *Treatment combinations:* Treatment loosely means subjecting some person or something to some action. In experimental design, it refers to a particular experimental manipulation or procedure under which participants are run. For example, in a 2 × 3 factorial experiment the participants are assigned 6 treatments. In the present text, the term treatment and treatment combinations will be used interchangeably. In a single-factor experiment, the levels of factor themselves constitute treatments. For example, suppose the experimenter is studying the impact of the level of illumination upon reading ability and he decides to have four levels of illumination. Here there will be four treatments and each participant will be randomly assigned to each of these three treatments if a randomized group design is followed. In multiple-factorial experiment such as 3 × 3 × 4, there will be 36 treatment combinations and each participant will be randomly assigned to one of these 36 treatment combinations.

(d) *Dimensions:* The dimensions of factorial experiment are indexed by the number of levels of each factor and the number of factors. For example, factorial experiment in which there are four factors, first having 2 levels, second having 3 levels, third having 3 levels and four having four levels, is called as 2 × 3 × 3 × 4 (read as two by three by three by four) factorial experiment. The dimension of this experiment is 2 × 3 × 3 × 4.

(e) *Main effects:* Main effects refer to the difference in performance from one level to another for a particular factor, averaged over the other factors. In factorial experiment, the mean squares (MS) for the levels of factors are known as the *main effect* of the factor. Let us illustrate the concept of main effects through an example. Suppose the experimenter has planned a $2 \times 3 \times 4 \times 5$ factorial experiment in which factor *A* has two levels, factor *B* has three levels, factor *C* has four levels and factor *D* has five levels. The sum of squares for *A* corresponds to a comparison between levels A_1 and A_2, the B sum of squares to a comparison between levels of B_1, B_2 and B_3, the *C* sum of squares to a comparison between levels of C_1, C_2, C_3 and C_4 and the *D* sum of squares to a comparison between D_1, D_2, D_3, D_4 and D_5. The difference in performance between levels of A_1 and A_2, averaged over the levels of *B*, *C* and *D* is called as the main effect of *A*. This is also called as A-effect. Likewise, the difference in performance among levels, B_1, B_2 and B_3, averaged over the levels of factors *A*, *C* and *D* is called main effect of *B*, and so on. This is also called as *B-effect*.

When graphically shown, the main effect is the curve joining points representing mean performance on the levels of a particular factors averaged over other factors of the experiment. A significant main effect will have significant slope, that is, the curve will not be parallel to *x*-axis.

(f) *Simple effects:* In any factorial experiment, the effect of a treatment on one factor at a given level of the other factor is known as the simple effect. Let us take an example of a 2×2 factorial experiment in which both factors *A* and *B* have two levels each ($A_1 A_2$, $B_1 B_2$). The effect of treatment on two levels of *A* under each of the two levels of *B* is called simple effect of *A*. Likewise, the effect of treatment on two levels of *B* under each of the two levels of *A* is called simple effect of B.

When plotted graphically, the simple effects resemble the two factor interaction. For example, the simple effect of *A* is the profile where the levels of the factor *A* are marked on *x*-axis and the two curves represented by the levels of B_1 and B_2 are the simple effects at each of the two levels. Likewise, the simple effect of *B* is the profile where the levels of factor *B* are marked on *x*-axis and the two curves represented by A_1 and A_2 are the simple effects at each level.

(g) *Interaction effect:* Interaction effect is an important effect in any factorial design because it allows the experimenter to investigate the interaction among the independent variables. An effect of two independent variables operating simultaneously and in combination on a dependent variable is called as interaction effect. In fact, it is a larger effect than the effect that occurs from the sum of each independent variable working separately.

Interaction is said to have occurred when change in one value of one variable brings a change in the effects of the other. An interaction effect is an effect in which the impact of one variable (on the measured variable) is different across the levels of other variable (Aron, Aron & coups, 2006). In simple words, there is an interaction between two factors if the effect of one factor depends on the levels of the second factor. When two factors are shown as *A* and *B*, the interaction is identified as the $A \times B$ interaction. Let us take an example to illustrate the interaction effect.

Suppose the experimenter tested a group of some rats ($N = 25$) on maze learning task. Some rats were administered drug at two months of age and some other rats were administered drug at the age of 10 months. There was also a group of rats that were administered no-drug at either 2 months of age or 10 months of age. Now the question is: Does this drug effect (one factor) depend upon age (the level of second factor)? The hypothetical data are presented below for explaining interaction.

A

	2 month	10 month	Overall
Drug	$\overline{X} = 70$	$\overline{X} = 90$	$\overline{X} = 80$
No-drug	$\overline{X} = 40$	$\overline{X} = 60$	$\overline{X} = 50$
Difference	30	30	

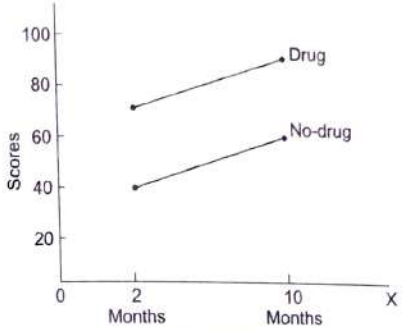

Fig. 21.1

B

	2 month	10 month	Overall
Drug	$\overline{X} = 50$	$\overline{X} = 110$	$\overline{X} = 80$
No-drug	$\overline{X} = 50$	$\overline{X} = 70$	$\overline{X} = 60$
Difference	0	40	

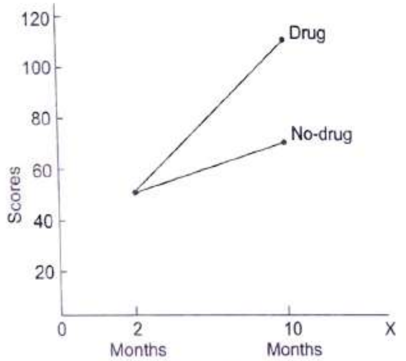

Fig. 21.2

C

	2 month	10 month	Overall
Drug	$\overline{X} = 70$	$\overline{X} = 40$	$\overline{X} = 55$
No-drug	$\overline{X} = 40$	$\overline{X} = 70$	$\overline{X} = 55$
Difference	30	30	

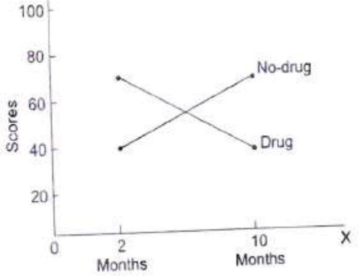

Fig. 21.3

If you look at data and graph presented in A, it will be obvious that such data presents no interaction. If you pay attention to main effect for the drug in this set of data, the overall average mean score for the drug group is 80 and for no-drug group is 50. This means that drug has a 30-point (80–50) effect. Does this drug effect depends upon age (levels of the second factor)? Now it is time to look at each group separately. For animals tested at 2 months, the drug effect is exactly 30 points (70–40) and for animals tested at 10 months, the drug effect is still 30 points (90–60). It means that drug effect does not depend upon age. Therefore, there is no interaction. When the effect of one treatment remains constant across all levels of other factor (showing no interaction), there will be a constant distance between two lines in the graph. Graph of such data will have *parallel lines.*

Now look at the data and graph presented in B. Again the overall drug effect is 20 points (that is, $\overline{X} = 80$ for drug group and $\overline{X} = 60$ for no-drug group of animals). But when you look at each age separately, you find a 40-point drug effect for animals given drug at 10 months of age, whereas such drug effect was zero at 2 months of age. It means that the effect of drug does depend upon the age of animals and therefore, there is an interaction. Here lines of the graph are not parallel rather they *converge* at a point.

Now look at the data and graph presented in C. For these data, the overall drug effect is zero (55–55). Both the drug and no-drug groups show an overall mean of 55. Does this mean that the drug has *no* effect? The answer is *no.* The drug produces 30 point increase in mean score for young animals and 30 points decrease in the mean score for old animals. Thus the effect of the drug depends upon which age-group is being examined. It means that there is an interaction. Lines in the graph *cross* each other providing evidence for interaction.

One point of caution is that when there is an interaction effect, the main effects may be misleading. Therefore, whenever an interaction is present, one must interpret the effects of each factor in terms of its interaction with the other factor.

SOME IMPORTANT TYPES OF RESEARCH DESIGN

As stated earlier, the design of an experiment is the blueprint of the procedures that enables the experimenter to test various hypotheses. It is, therefore, essential that the experimenter at the outset must decide what type of design he will frame for testing the hypotheses. The selection of a design is partly dependent upon the experimenter's decision regarding whether he is going to use one or more than one group of subjects in the proposed experiment. If he decides that there will be only one group of subjects who will be tested under different values (or conditions or treatments) of the independent variable, the resulting experimental design is known as the *within-subjects design* or *repeated treatment designs.* If, on the other hand, he decides to use a separate group for each value of the independent variable, the resulting design is referred to as the *between-subjects designs* (Kantowitz & Roediger 1984). Thus based upon the criterion of the number of groups used in the experiment, there are two types of research designs—*between-subjects design* and *within-subjects design.*

In psychological and educational researches, the between-subjects design has been used more frequently than the within-subjects design. Before we go into the details of these two types of designs, we shall discuss the main subtypes of these designs. The between-subjects design is divided into three most common types: *randomized-groups design, matched-groups design* and *factorial design.* A randomized-groups design is one in which subjects are randomly assigned to the different groups meant for the different conditions or values of the independent variable. It is based upon the assumption that the random assignment of subjects into two or more groups will make these groups statistically equivalent on the subject relevant variables (attitude, ability, capacity, etc.), which may produce variations in the dependent variable. When the subjects are randomly assigned to only two groups, the resulting experimental design is known as two-randomized-groups *design* and when the subjects are randomly assigned to more than two groups, the resulting design is known as the multi-groups design or more-than-two-randomized-groups *design.* Matched-groups *design* is a type of between-subjects design in which the subjects are matched depending upon mean, standard deviation, pairs, etc. Likewise, a *factorial design,* which is also a type of between-groups design, may be defined as the design in which two or more than two independent variables are studied in all possible combinations (each combination having a separate group of equal or unequal subjects) so that their independent and interactive effects may be studied on the dependent variable.

Within-subjects design is of two types—*complete within-subjects design* and *incomplete within-subjects design.* A complete within-subjects design is one in which practice effects are balanced by administering the conditions several times to each subject, in different orders each time, such that the results for each subject become interpretable. An incomplete within-subjects design is one in which each condition is administered to each subject only once while varying the order of administering the conditions across subjects in such a way that practice effects can be neutralized when the results for all subjects are combined. The common techniques for balancing practice effects in complete within-subjects design are block randomization and ABBA counterbalancing while in incomplete within-subjects design, one common technique is to use all possible orders of the treatments and subsequent assignment of each subject to one of the orders.

For easy understanding of the above designs with the entire picture at a glance, Figure 21.1 has been given which arranges them into a schematic presentation.

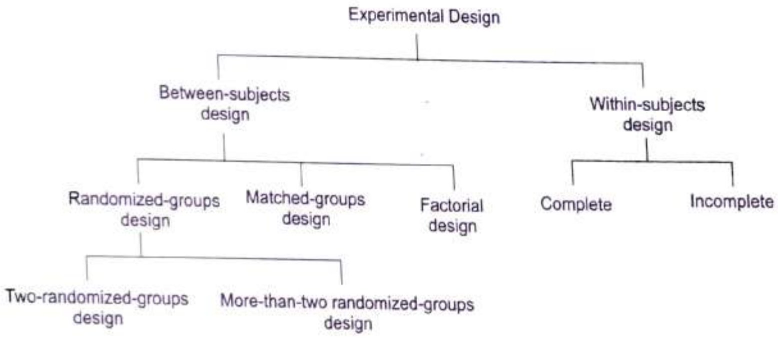

Fig. 21.1 Schematic representation of experimental design

A critical evaluation of each of these designs is presented below.

BETWEEN-SUBJECTS DESIGN

Two-Randomized-Groups Design

A two-randomized-groups design is so called because here the subjects are randomly assigned to two groups only. In formulating this type of design, the experimenter first defines the independent and the dependent variables. Subsequently, he selects two values of the independent variable. These two values may also be called 'conditions' or 'treatments' of the experiment. His main interest is to examine whether or not these two conditions affect the dependent variable in a differential way. In selecting a sample for the proposed experiment, he defines the population. Suppose the experimenter wants to study the effect of reward upon the rate of learning verbal concepts among kindergarten children. For this, he specifies the population which consists of all kindergarten children in a particular district or state. He wants to make a generalization about this population. Further, suppose that this population consists of 5,000 children in a district. He may randomly select a sample of 100 children. These subjects may be regarded as representatives of the population. In this experiment the experimenter may wish to have two values or conditions of reward—presence of the reward or rewarded condition, and absence of the reward or nonrewarded condition. Accordingly, the 100 subjects will be divided into two groups. The experimenter may randomly assign the subjects into two groups. For this, he may write the names of the subjects on separate slips of equal size and same colour, or he may number them on a separate slip. These 100 slips may be folded in like manner and placed in a box for a reshuffle. The experimenter may decide that the first slip will go to the first group, the second slip to the second group, the third slip to the first group, the fourth slip to the second group, and so on. This process will naturally result in two equal *N* groups of 50 each. Now, just by a toss of a coin, the experimenter may again decide which group will be named as the experimental group (rewarded group) and which group as the control group (nonrewarded group). It is expected that these two randomly assigned groups will not differ singnificantly at the start of the experiment. The experimental group is given one type of treatment and the control group is given another type of treatment. In the above example the children of the experimental group will be rewarded for learning the verbal concepts, whereas the children of the control group will not be rewarded for learning the same task. The nature of the reward will depend upon the predetermination of the experimenter—it may be verbal or monetary. Subsequently, the scores of all subjects of these two groups on the dependent variable (learning of verbal concepts) will be recorded and subjected to statistical analysis. Usually the *t* test or its nonparametric substitute, the Mann–Whitney *U* test, is applied in a two-randomized-groups design. (A detailed discussion of these statistical tests is done in Chapter 23.) The experiment may be continued for several days or weeks. If the statistical test reveals that these two groups significantly differ on the measure of the dependent variable, it is concluded that the difference in the dependent variable is due to the experimental manipulation of the independent variable. If, for example, the experimental group is 5 units ahead in learning the verbal concepts and this difference is statistically significant, we conclude that reward accelerates the learning of verbal concepts in kindergarten children.

Before concluding the discussion of the two-randomized-groups design, a retreat is essential here. As it is obvious from the above interpretation, the heart of the two-randomized-groups design is a random assignment of subjects into groups so that there will not be any systematic relationship between any of the characteristics of the subjects and the particular group to which they are radomly assigned. The question is how to achieve this end. Underwood (1966) has suggested two primary ways through which unbiased groups or random groups of subjects can be formed: *captive assignment and sequential assignment.*

The technique of captive assignment is one in which all subjects are individually known to the experimenter by name and they are all present at one time. In a way, they are made captive for the entire duration of the experiment so that they can at random be assigned to the different conditions or groups. That is why the technique is known as a captive assignment. When the experimenter is to use all subjects from a single class, this best illustrates the situation prevailing in the captive assignment. Several submethods can be adapted for randomly assigning the captive subjects into different groups. Suppose there are 50 students in a class. The experimenter wants to conduct a two-condition experiment. He may, then, randomly assign these subjects into two groups of preferably equal *N*. It is not necessary that in a randomized-groups design, the number of subjects in each group should be equal. But the equal *N* is preferred because it facilitates some statistical calculations (Edwards 1968; Winer 1971). For the process of random assignment of subjects into different groups, any of the following procedures can be adopted.

1. The most common method of randomly assigning the subjects into the different groups is to use the Table of Random Numbers. The use of the Table of Random Numbers in selecting the subjects may be illustrated with an example. Suppose 50 students are to be randomly assigned to two groups. Their names are alphabetically arranged. Any convenient starting point in the Table of Random Numbers is chosen and we go on attending to the number if it is either 1 or 2. Suppose the sequence of random numbers is 3628213. Then only 221 is selected and the remaining numbers are ignored. Now, the first student in the alphabetical list is assigned to the second group, the second student is also assigned to the second group and third student to the first group, and so on, until both the numbers have occurred 25 times. It may be that any one number (that is, either 1 or 2) may complete 25 frequencies earlier than the other. In such a situation, that number in further sequence is ignored so that each group may consist of 25 subjects.

2. Each of the 50 students may be assigned a number from 1 to 50 on separate slips of equal size and of the same colour. Subsequently, the slips may be folded in a similar way and kept in a box. After proper reshuffling, the numbers are drawn one by one by the experimenter. The first 25 numbers drawn may constitute the first group and the remaining 25 numbers drawn may constitute the second group.

3. Suppose the students are allowed to take a seat in the classroom. When all the 50 students have taken their seats according to their choice, the experimenter may start from the left (or right) row and go on counting until he has completed 25. This will constitute the first group. The remaining students may be counted off in a similar way to constitute the second group. This procedure of assignment of subjects into groups is not a completely random or unbiased one because it may produce a systematic relationship or correlation between 'the choice for sitting in the class' (a subject variable) and the group to which it is assigned.

4. The simplest way of randomly assigning the subjects into various groups is to place the subjects in the list in alphabetical order. Subsequently, the first student in the list is assigned to the first group; the second student to the second group; the third student to the first group, and so on.

The technique of sequential assignment is another method of random assignment of subjects to different groups. In this technique, the experimenter does not know the details of the subjects. He is simply aware of the fact that a certain number of subjects will participate in the experiment. As the experiment in the randomized-group design usually continues for several days or weeks, the experimenter may run five subjects on the first day, two subjects on the second day, six subjects on the third day, and so on. Thus the experiment is conducted in a prearranged sequence. There are two main techniques of sequential assignment of subjects into the various groups, namely, *complete randomization* and *block randomization*.

The method of complete randomization as applied to the technique of captive assignment can also be extended to the technique of sequential assignment. Let us suppose that a total pool of 50 subjects is available to the experimenter for conducting a five-condition experiment, each condition having 10 subjects. Let us name the five conditions of the experiment as A,B,C,D and E. The simple way is to prepare 50 slips of the same size and colour and number them from 1 to 50 so that each slip has only one particular number. Of these 50 slips, 10 will have 'A' written or printed on them, 10 will have 'B' written or printed, 10 will have 'C', 10 will have 'D' and the remaining 10 will have 'E'. Subsequently, these 50 slips will be kept in a box. After stirring, the experimenter draws the slips one by one and lists them in the order drawn. Then, the first subject coming to the experimenter is assigned to the condition written on the first slip drawn, the second subject is assigned to the condition written on the second slip drawn, and so on. Suppose the first slip drawn contains C and the second slip contains D. It means that the first subject coming to the experimenter will be assigned to condition C, and the second subject coming to the experimenter will be assigned to condition D. By doing so the experimenter is obviously trying to avoid the relationship or correlation (which may bias the assignment) between the particular condition, and the time when the subjects come to the laboratory. One general limitation of this technique is that sometimes, just by chance, there may occur some correlation between the assigned condition and the times when subjects appear. Suppose the 50 drawn sequences beome like this: CCCCACCCDCCC...EBBAABBBABBBB. If this type of sequence emerges, it is obvious that till the formation of the C condition group (of 10 subjects), B condition does not appear. B entries become very frequent towards the end of the series. This means that there is some correlation, though by chance, between B condition and the time succession. When the number of subjects is 30 or more in each group, such a type of sequence is very unlikely. But when *N* is 10 or less than

10, the probability for such a sequence is high. The remedy for such a biased sequence is to use the technique of block randomization.

The method of block randomization may be defined as one in which each condition or value of the independent variable occurs once in each successive block of trials, but the order of conditions within a block is random and different from every other condition. To meet the requirement of block randomization, it is essential that all conditions must occur an equal number of times. The number of times the conditions should occur is determined by the number of blocks. If, for example, there are six blocks, each condition will occur six times altogether. In preparing a schedule of conditions representing block randomization, the table of random numbers may be used. As an illustration of block randomization of five conditions of the above experiment, the following sequence of schedule may be followed if each condition is to be presented 10 times:

CADBE; BADEC; CAEDB; BACED; ADCBE; EDBAC; ADECB; BEACD; DABCE; CDBAE

In the above preparation for block randomization, there are ten blocks and each condition occurs within each block once in a random order. In the above sequence, the order indicates the pairing of a condition with successively appearing subjects. In other words, the first subject will be assigned to condition C; the second subject to condition A; the third subject to condition D; the fourth subject to condition B; the fifth subject to condition E; the sixth subject to condition B, and so on, and the 50th subject to condition E.

After the experiment is over, the statistical analysis of the data is done. In a two-randomized-groups design, generally the *t* test or the Mann–Whitney *U* test is the appropriate statistic to follow. The details of calculation of these statistics in this type of design appear in Chapter 23.

The two-randomized-groups design has two important *limitations. First,* in the course of running an experiment based upon the two-randomized-groups design (or more-than-two-randomized-groups design) usually the experimenter loses some subjects. The problem posed by the differential loss of subjects after their random assignment to groups is serious and relates to the fact whether or not subjects were true representatives of the original random or unbiased group. If the subjects are not representative of the original randomly assigned groups, it is obvious that these equivalent groups resulting from random assignment may no longer be really equivalent and then the conclusion drawn may not be generalized. Let us illustrate this point with an example. Suppose the experimenter wants to run a two-condition experiment. In one condition 20 trials per day up to six days are to be given, and in another condition 200 trials per day up to six days are to be given. A total of 40 subjects are randomly assigned to two groups ($N = 20$ in each), one meant for the 20-trials-per-day condition and another meant for the 200-trials-per-day condition. Further suppose that five of the original 20 subjects randomly assigned to the 200-trials-per-day condition do not turn up on the 4th, 5th and 6th day of the experiment, whereas all subjects assigned to the 20-trials-per-day condition turn up for all the six days. The performance of these two groups can be compared through appropriate statistical tests despite unequal *N*, but the question arises as to why the experimenter has to face the loss of five subjects in one group. Was the task boring? Were the subjects not interested? Were they giving least importance to the experimental task? Did they find it convenient to become ill? Whatever the reasons may be, it is obvious that such subjects are not representative of the original unbiased group resulting from the random assignment and in such a situation, these two groups formed from random assignment and said to be equivalent, remain no longer equivalent ones.

Second, behavioural researches have two general purposes: one is to determine which of the many independent variables tends to influence the dependent variable most and another is to determine what type of relationship exists between the influential independent variable and the dependent variable. A two-randomized-groups design serves the first purpose but does not serve

the second one. Since in such a design only two values of independent variables take part in experimentation, it is difficult to state precisely the exact relationship between the influential independent variable and the dependent variable. For stating such a relationship, it is essential that more than two values of the independent variable are studied. Such a design is known as more-than-two-randomized-groups design whose discussion in detail is presented below.

More-Than-Two-Randomized-Groups Design

Because of some limitations of the two-randomized-groups design, it is being used in a restricted way and more emphasis is being given to the more-than-two-randomized-groups design. Such a design is also known as the *multi-groups design.* As its name implies, in such a design there are three or more conditions or values of the independent variable and accordingly, three or more groups of subjects participate in the experiment. The design is so called because the total subjects are randomly assigned to three or more unbiased groups. The process of captive assignment as well as that of sequential assignment (discussed in the previous section) may be utilized as the techniques of random assignment of subjects into three or more groups.

In psychological and educational researches, the use of the more-than-two-randomized-groups design is more common than the two-randomized-groups because the former has three distinct advantages over the latter.

1. Suppose that an experimenter wants to know which of the four methods of teaching Russian (say, Method A, B, C, and D) is to be preferred. Suppose 80 students of the same age, intelligence and sex are available for this purpose. The researcher's first step would be to assign them randomly into four groups, each group having 20 subjects. These groups are supposed to be equivalent groups after random assignment. Subsequently, one group will be taught by method A; the second group by method B; the third group by method C; and the fourth group by method D. All subjects will be administered a test of Russian and through the appropriate statistical techniques, we can answer the above question. This illustrates the more-than-two-randomized-groups design where all the four groups are used simultaneously. It is also possible to answer the question by conducting six separate two-groups experiments. For example, in one experiment subjects taught by method A would be compared with subjects taught by method B; in the second experiment, subjects taught by method A would be compared with subjects taught by method C; in a third experiment subjects taught by method A would be compared with subjects taught by method D; in a fourth experiment, subjects taught by method B would be compared with subjects taught by method C; in a fifth experiment, subjects taught by method B would be compared with subjects taught by method D; and finally, in the sixth experiment subjects taught by method C would be compared with subjects taught by method D. Needless to say, this type of experimental design is painstaking, time-consuming and strenuous. Not only this, the problem of controlling relevant variables will also be serious and manifold. Even the experimenter may behave differently in the first experiment and in the last experiment due to fatigue or boredom. Therefore, we can say that the multi-groups design or the more-than-two-randomized-groups design is superior to the two-randomized-groups design.

2. In behavioural research it is, in general, regarded as a healthy practice to sample more values of the independent variable because this helps the investigator to evaluate better the influence of the independent variable upon the dependent variable. This purpose is served more satisfactorily by the multi-groups design than by the two-groups design. An example may illustrate this point clearly. Suppose the experimenter is interested in testing the hypothesis that the higher the hunger drive, the more correct responses a rat would make in a maze. For conducting an experiment of this type with multi-groups designs, he may take five groups of rats such that Group I has zero hour of food deprivation; Group II has three hours of deprivation; Group III has six hours of deprivation; Group IV has eight hours of deprivation; and Group V has

ten hours of deprivation. If the experimenter wishes to conduct a two-groups design, the immediate problem is which of the five values of the independent value would tend to show the experimenter's choice? It is expected that any sensible experimenter would like to choose zero hours of deprivation (which would act like the control group) for one group and the second group might be of a six-hour duration. Now, suppose that the result of the experiment based upon the multi-groups design was that there was no significant mean difference in the correct responses of the zero-hours group, three-hours group, six-hours group, and eight-hours group, but the ten- hours group was better than the first four groups because it produced a significantly higher mean (of the correct responses) than the first four groups. The conclusion would be that the greater hours of deprivation lead to more correct responses, although 0 hours to 8 hours of deprivation produces no significant difference in the means of the correct responses. When the two-groups design (0-hours and 6-hours deprivation groups) are used to study this problem, the most obvious conclusion would be that the variations in the hours of deprivation does not affect the means of the correct responses in the maze. This conclusion would then clearly be wrong. The reason for drawing a wrong conclusion is that only two values of an independent variable were sampled. In the multi-groups design, the possibility of drawing this type of conclusion is automatically minimized.

3. As stated in the discussion of the two-randomized-groups design, such a design fails to establish an adequate relationship between the influential independent variable and the dependent variable, probably because this type of relationship can be established only after several values of the independent variables are sampled and averaged. In a multi-groups design, it is easier to establish adequate relationship between the independent variable and the dependent variable because the design utilizes several conditions or values of the independent variable. This point can be illustrated with the help of Figures 21.2 and 21.3. In Figure 21.2, the *x*-axis shows the increasing values of the independent variable and the *y*-axis shows the increasing values of the dependent variable. Group I was given two hours of practice on memorizing a list of nonsense syllables, and Group II was given eight hours of practice for the same task. It is obvious that the value of the dependent variable (or proportion of retention of the nonsense syllables) for Group I is less than the proportion or value of Group II. As this design tells nothing about what happens in between these two limits (two hours of practice and eight hours of practice), we are unable to conclude that any obvious relationship exists between the independent variable (hours of practice) and the dependent variable (retention of the task). Suppose we add two more groups so that the four groups each having equal hours of practice become: 2 hours (Group I), 4 hours (Group II), 6 hours (Group III) and 8 hours (Group IV). The resulting data from this experiment have been plotted in Figure 21.3. It is obvious that each successive group has obtained a higher value on the dependent variable, that is, Group II obtains a higher value than Group I; Group III obtains a higher value than Group II; and Group IV obtains the highest value on the dependent variable. From this type of data, it appears that the four values of the independent variable are related in a certain way with four values of the dependent variables. Obviously, the relationship is such that it can be easily explained with a straight line and hence, we can conclude that a linear relationship exists. Of course, there are other types of relationships, which can also be demonstrated. A frequently reported relationship is also a nonlinear or curvilinear relationship between the dependent variable and the independent variable.

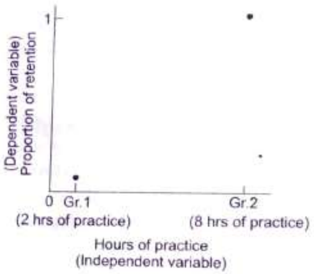

Fig. 21.2 Unknown relationship between dependent variable and independent variable obtained from an experiment based upon a two-groups design

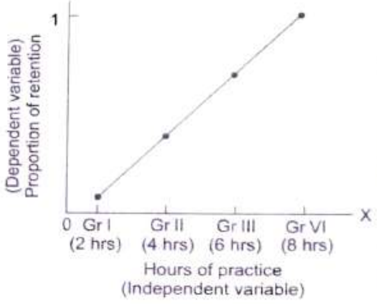

Fig. 21.3 A linear relationship between independent variable and dependent variable obtained from an experiment based upon a two-groups design

To summarize, the multi-groups design enables the experimenter to establish an adequate relationship between the dependent variable and the independent variable but the two-groups design lacks this trait.

In a multi-groups design, the two most common statistics applied are the analysis of variance and the Duncan Range test. For a detailed discussion of these statistics, particularly Duncan Range test, readers are referred to Downie & Heath (1970) and Siegel (1956).

Matched-Groups Design

Like the randomized-groups design, the matched-groups design (also known as the randomized-block design) may be a two-matched-groups design or a more-than-two-matched-groups design. Whatever the type, in the matched-groups design all subjects are first tested on a common task or a pretest measure (also called the matching variable) and then, they are formed into groups (as many as needed for the experiment) on the basis of the performance of the common task or the matching variable. The groups thus formed are said to be equivalent groups. Subsequently, the different conditions or values of the independent variable are introduced to each group. If these groups have equivalent means on the dependent variable before the experimental treatment is given and if a significant difference occurs after administering the experimental treatment and controlling the relevant variables, the resulting differences in the dependent variable may safely be attributed to the experimental treatment. Thus the matched-groups design is simply a way of establishing the fact that all the groups have equal dependent variable values prior to the administration of the experimental treatment.

The matched-groups design is based upon the principle that the experimental unit (or subject) can form a block or group. As a matter of fact, a group of subjects said to be homogeneous with respect to the matching variable forms a block. It is expected that each block of subjects will be equivalent or homogeneous on the dependent variable in the absence of the experimental treatment than subjects selected at random. In fact, each block of subjects is matched with respect to a matching variable. This is why the matching group design is also known as the randomized block design (Edwards 1968, 156), a term which has been borrowed from the field of agriculture in which the experimental unit is a plot of land and a block is defined as the strip of several adjacent plots. It is expected that plots which are adjacent to each other will be more homogeneous in fertility and other soil characteristics than equal numbers of plots selected at random from different fields. In psychological and educational researches, the experimental unit corresponding to a plot in agricultural research is the subject.

Now, two related problems arise in connection with the matching task: one is concerned with the selection of the matching variable and another with the ways or *methods* of matching.

Selection of the Matching Variable

As discussed above, in the matched-groups design subjects are measured on the basis of the matching variable prior to the introduction of the experimental variable. The question is: How is one to select a matching variable? The most important characteristic on the basis of which a matching variable can be selected is its ability to yield a high correlation with the experimental task or the dependent variable. If the matching variable is found to be highly correlated with the scores on experimental task or with scores on the dependent variable, the matching may be regarded as successful. On the other hand, if the matching variable yields a poor correlation or no correlation with the dependent variable scores, the matching is not regarded as successful. Now, the question arises: How can the experimenter find a matching variable which highly correlates with the dependent variable? One obvious answer is to use the dependent variable itself as the matching variable. For example, suppose the experimenter is studying the effect of knowledge of results upon the maze learning. There may be two groups or subjects for it, one group working with knowledge of results in the trials given to it and another group working without knowledge of results in the trials given to it. For assigning the subjects into two groups by the matching technique, the experimenter may give them six trials on the maze learning itself and obtain a set of scores. Subsequently, based upon the obtained scores the subjects may be paired off.

The method used for dividing each pair into two groups is randomization. The random assignment of subjects is essential to prevent the operation of the experimenter's bias in matching. Then, these two groups may be given trials according to the experimental plan. The *t* test may be calculated between two sets of scores obtained by these two matched-groups as a measure of whether or not the knowledge of results has produced a significant mean difference. One biggest advantage of using the initial trial performance score on the dependent variable is that the correlation between the matching variable and the dependent variable is undisputed, and hence, the success of matching is almost certain. But the experimenter may have to face a situation where initial trials on the dependent variable may not be used as the matching variable. In such a situation he will have to depend upon some other measure, which should be different from the dependent variable but highly correlated with it. Suppose that the experimenter is studying the rapidity (and not the effect of knowledge of results) involved in the process of tracing a stylus maze by a human subject. In such a situation if the experimenter decides to match the subjects on the basis of scores of six trials given in maze tracing, this is likely to give some extra benefit to the subjects in tracing of the maze when it is introduced as an experimental task (or the dependent variable because such a practice may acccelerate the rapidity in tracing the maze). Likewise, if the experimenter is interested in studying the effect of practice in numerical ability, he cannot match subjects on the basis of the initial scores obtained from the solution of the arithmetical problems by the subjects in five or six trials because in that case each subject would be knowing the answer of the numerical problems when they (the arithmetical problems) would be used as the experimental task. However, in such a situation there is one solution. The experimenter may use different numerical problems (than those used as dependent variable) as a matching variable. Or, if the test of numerical ability has a large number of problems, half of them may be used as matching variables and the remaining half may be used as the measure of the dependent variable. In either solutions suggested above, the matching variable is expected to yield a high correlation with the dependent variable .

In matching, an independent task may be used as the matching variable but the relationship between the independent task and the dependent variable should have been established in pevious researches or investigations. It may, therefore, be said that a matched-groups design should not be used in a situation where the matching variable does not yield or expect to yield a high correlation with the dependent variable.

Methods of Matching

Having selected a matching variable and obtained a set of scores earned by all subjects on the matching variable, the next step is to match them. There are two ways of matching: *matching by pairs* and *matching on the basis of mean and standard deviation.*

Matching by Pairs: One very convenient technique is matching by pairs. On the basis of the obtained scores by the subjects, the experimenter matches subjects in a way that each subject has a *corresponding* partner in the matched group or groups. An example may be given to illustrate matching by pairs. Suppose the experimenter wants to make a comparative study of the lecture method and the demonstration method upon problem-solving behaviour. There are 14 subjects available for the purpose. If the experimenter is to use the randomized-groups design, he may randomly assign 14 subjects into two groups, irrespective of what he knows about the subjects. But in using the matched-groups design, he must, first, match them on the matching variable. Let us suppose that intelligence test scores are used as the matching variable. All subjects are administered the intelligence test and their scores are obtained, which have been shown in Table 21.1. Now, the experimenter wants to construct two such groups (one for the lecture method and another for the demonstration method), which are equal on the intelligence test scores. For this, the experimenter chooses subjects who have equal scores. Thus subject no. 1 is paired with subject no. 9; subject no. 2 is paired with 12, and so on, and in this way, seven pairs are formed.

Table 21.1 Scores of 14 subjects on intelligence test (data are fictitious)

Subjects	Scores	Subjects	Scores
1	70	8	90
2	75	9	70
3	69	10	81
4	90	11	85
5	69	12	75
6	81	13	62
7	85	14	62

Table 21.2 Matched-groups formed on the basis of the scores of the matching variable

Lecture method group		Demonstration method group	
Subject No.	Intelligence score	Subject No.	Intelligence score
1	70	9	70
2	75	12	75
5	69	3	69
8	90	4	90
6	81	10	81
11	85	7	85
14	62	13	62

Each pair, thus formed, constitutes a block. From these blocks, or pairmates, two groups are randomly formed. By a simple toss of a coin, the experimenter may determine that subject no. 1 goes to the group to be taught by the lecture method and subject no. 9 goes to the group to be taught by the demonstration method. Sometimes, it has been reported that the assignment of subjects is completely left to the will of the experimenter. This is a dangerous practice because the element of bias may be introduced in the assignment. For example, the experimenter may place all those subjects who are highly motivated and interested in the group to be taught by the demonstration method. But when random assignment is done, the experimenter's bias is automatically controlled. The two-matched groups formed on the basis of intelligence test scores have been shown in Table 21.2. Subsequently, both groups are given the problems for solution and the score obtained by them is written as shown in Table 21.3. A *t* test may be applied to determine whether or not the obtained mean difference between the two sets of scores is significant. The method of calculation of *t* in the two-matched groups design will be discussed in Chapter 23.

Table 21.3 Scores on a problem-solving task by pairs of subjects matched on the basis of intelligence test scores

Lecture method group		Demonstration method group	
Subject No.	Intelligence score	Subject No.	Intelligence score
8	120	4	123
11	109	7	110
6	95	10	99
2	80	12	87
1	91	9	80
5	75	3	72
14	70	13	66

Computationally simpler than the *t* test is the *A*-test which may also be used in the two-matched-groups design as a substitute for the *t* test (Sandler 1955). Note that for facilitation in statistical calculation the subjects have been ranked in terms of scores on the dependent variable, the pair securing the highest scores on the problem-solving task being kept at the top and the pair securing the lowest score on the problem-solving task being kept at the bottom.

There is, however, one limitation of the technique of matching by pair. If the experimenter insists on precise matching, the persons with deviant scores on the matching variable are likely to be eliminated. Say, for example, that subject no. 1 and 9 are paired because each has an equal score, that is, 70. Suppose, for the time being, that any one of this pair has 65 instead of 70. Now in such a situation, these two subjects cannot be paired. Not only this, both the subjects would be dropped from the original group because they have deviant scores in the sense that there are no matching partners for them in the group. A situation like this is not uncommon when the experimenter is dealing with a larger group. This has a natural effect of shortening the original group. The problem created by this situation is, however, not serious unless the experimenter is faced with the task of very accurate generalization regarding the population from which the sample was drawn. As a partial solution of the problem, the experimenter may predetermine certain allowances. For example, he may predetermine an allowance for the difference of 5 scores and then, subject nos. 1 and 9 may be paired together despite the difference of the 5 scores on the matching variable between them.

Matching in Terms of Mean and SD: Another way of matching the subjects is in terms of central tendencies and variabilities in the distribution of scores obtained on the matching variable. Mean and *SD* are commonly selected as the measures of central tendency and variability of the distribution respectively. When the basis of matching is the mean and the standard deviation of the scores obtained on the matching variable, the experimenter forms as many groups as needed for the experiment in a way that these means and standard deviations do not statistically differ. There are three methods of matching in terms of means and standard deviations, namely, the *random-blocks methods*, the *method of counterbalancing order* and the *block-repetition method*.

In the random-blocks method the blocks are, first, formed. The number of subjects in each block is kept constant and is determined by the number of values of the independent variable. Subsequently, subjects from each block are randomly assigned to the different groups as required by the different conditions or values of the independent variable. Suppose the experimenter wants to conduct a three-conditions experiment. Naturally, then, the number of values of the independent variable would be three. So, there would be three groups of subjects, one for each condition. Let us name those groups as A, B and C, and assume that there are 24 subjects available for the purpose. Since there are three values or conditions of the independent variable, there would be three subjects per block. Thus the maximum of eight blocks will be prepared for 24 subjects. The scores of 24 subjects on a matching variable (say, intelligence test) are presented in Table 21.4. They have been shown arranged in decreasing order to yield facilitation in using the above three techniques of matching. In using the technique of the random-blocks method, the experimenter may start from the top or bottom of the distribution. Let us start from the top. The first three subjects would form one block. Then these three subjects would be randomly assigned to groups A, B and C, with one subject to one group. Suppose the sequence of the random assignments is 2-1-3. The remaining subjects would be treated in the same fashion, that is, sequentially blocking them into groups of three subjects and subsequently, assigning them randomly to each group in a manner that one subject is assigned to one group. The random order in the first block is 2-1-3 (1, 2, 3 stand for conditions A, B, C respectively), that is, the first subject

Table 21.4 Scores for 24 subjects obtained on the matching variable and presented in decreasing order (data are fictitious)

Subject No.	Score	Subject No.	Score	Subject No.	Score
1	161	9	147	17	131
2	160	10	144	18	131
3	158	11	140	19	125
4	155	12	140	20	124
5	154	13	138	21	123
6	152	14	136	22	120
7	148	15	135	23	118
8	147	16	132	24	117

was placed in condition B; the second subject in condition A and the third subject in condition C. Another new random order would be determined for the next block and subjects would be, accordingly, assigned to different groups. The resulting data has been shown in Table 21.5.

In the method of counterbalancing the order of matching, namely ABCCBA...etc., a pattern would be adopted for assigning the subjects into three groups. Thus the first subject would be assigned to group A or condition A; the second to group B; the third to group C; the fourth to group C; the fifth to group B; the sixth to group A; and the seventh to group A, and so on. It is assumed that subjects thus assigned into the different groups would be matched groups. The resulting data has been shown in Table 21.5.

Table 21.5 Three groups A, B and C are matched in terms of means and SD with three different procedures on the basis of the data shown in Table 21.4

Random-blocks method			Method of counterbalancing order			Block-repetition method		
A	B	C	A	B	C	A	B	C
160	161	158	161	160	158	161	160	158
155	152	154	152	154	155	155	154	152
147	147	148	148	147	147	148	147	147
144	140	140	140	140	144	144	140	140
136	138	135	138	136	135	138	136	135
132	131	131	131	131	132	132	131	131
124	123	125	125	124	123	125	124	123
117	120	118	117	118	120	120	118	117
Means 139.375	139.000	138.625	139.000	138.750	139.250	140.375	138.750	137.875
SDs 13.93	13.23	13.15	13.60	13.55	13.17	13.41	13.55	13.25

In the block-repetition method (so named by the author because a block of the required conditions is first formed and then successively repeated), a block of three conditions (or required number of conditions) is made in the natural sequence and the same order of the sequence of conditions is repeated in each block. For example, ABC is one block, which appears in a natural sequence and the same order of sequence would be maintained in each of the eight blocks. Thus ABC ABC ABC... sequence would be repeated in each block. The sequence indicates that the first subject would go to group or condition A; the second to condition or group B; the third to condition or group C; the fourth to condition A, and so on. The resulting data has been shown in Table 21.5. The means and standard deviations for each column in Table 21.5. have been given at the bottom. Of the three methods, it is obvious that there occur greater mean differences in the block-repetition method and the least mean differences in the random-blocks method. It can, therefore, be recommended that the block-repetition method should be the least preferred method and the random-blocks method should be the most preferred method for matching.

Now, one question might be raised here. Which method of matching should the experimenter follow? Should he match on the basis of means and standard deviations or should he match by pairs? There is no fixed answer to this question. It all depends upon the experimenter's purpose—whether he wants statistical precision in matching or he can tolerate

some differences in ability levels of subjects. In the former case, he should adhere to matching by pairs and in the latter situation he should resort to matching on the basis of means and standard deviations.

Which Design to Prefer: Randomized-Groups or Matched-Groups?

If the experimenter has decided to use the independent groups in his experiment, should he use the randomized-groups design or the matched-groups design? Different experts would give different answers to this question which in itself is an evidence that there is no strong point in favour of choosing one over the other. A review of the literature clearly shows that the randomized-groups design has been used more frequently than the matched-groups design. There are two probable reasons for this. First, in the matched-groups design the inconvenience of the experimenter is increased because of two factors. One reason is that it is difficult to obtain a relevant matching variable and obtain a set of scores upon it. Another difficulty arises when the experimenter is using the initial trial performance score as the matching variable. If the problem is to be investigated within a laboratory, the experimenter will have to bring the subjects twice to the laboratory, the first time for obtaining initial scores and the second time for obtaining scores after introducing the independent variable. This requirement of keeping the subjects present twice in the laboratory is troublesome. For the reasons stated above, the experimenter should prefer the randomized-groups design to the randomized-block design or the matched-groups design. Second, many statistically advanced techniques can more easily be applied with the randomized-groups design than with the matched-groups design. So, the former has been preferred to the latter.

Apart from this, in the matched-groups design the experimenter has fewer degrees of freedom (df) than in the randomized-groups design if the total number of subjects (N) is equal in both the designs. For example, suppose the experimenter has 20 subjects. In the two-matched-groups design there will be 10 pairs of subjects and accordingly, the degree of freedom will be equal to $10-1=9$. But in the randomized-groups design, there would be $20-2=18$ degrees of freedom. If the experimenter is to use the t test as a measure of significance of the mean difference after the experimental treatment, it may be that the same value of t may be significant with the randomized-groups design but not significant with the matched-groups design probably because of the fact that the greater the number of degrees of freedom, the smaller the value of the t required for being significant. This fact may be illustrated with an example. Suppose the experimenter gets 2.20 as the value of the t test in an experiment having two groups, each consisting of 10 subjects. If the value of the t test is interpreted on the basis of df calculated in the randomized-groups, that is, $20-2=18$, we find that this may be taken as significant at 0.05 level. The required value of t at the said level for the said df is 2.10. But the same value of t (that is, 2.20) is not significant if it is interpreted on the basis of df calculated in the matched-groups design (that is, $10-2=8$) at 0.05 level because the required value of t at this df is 2.31. For this reason, the matched-groups design suffers a disadvantage as compared to the randomized-groups design.

If there is a positive and high correlation between the matching variable and the dependent variable, our matching would be most perfect and successful. In such a case, the groups will be highly homogeneous or equivalent and then, the probability is that a sharp distinction may occur between the groups after the experimental treatment is given. In such a situation, the value of t is likely to be high so that it can refute the null hypothesis at a higher level of confidence. In a nutshell, it can be said that if there is a higher and positive correlation between the matching and the dependent variable, the value of t may be increased. And if the correlation is high, the experimenter can afford to lose some degrees of freedom by matching and then, he should prefer the matched-groups design to the randomized-groups design. If in reality this is not so, the experimenter should prefer the randomized-groups design to the matched-groups design. It may also be said that the loss of *df* consideration is a reality, which goes against the matched-groups design only when *N* is small. If the *N* is large, the experimenter would not lose much by matching subjects (even in the case of a small correlation between the matching and the dependent variable) because there is a very small difference between the values of *t* required at different levels of significance with greater *df*. With a larger *N* (and therefore, with higher *df*) the matched-groups design will yield significant statistical results (like the *t* test), similar to those found by the randomized-groups design.

Factorial Design

So far we have discussed the randomized-groups design and the randomized-blocks or the matched-groups design, which are appropriate for studying a single independent variable at a time. If the single independent variable is varied in two ways (and thus there are two values of the independent variable), the two-randomized groups or the two-matched-groups design is recommended and when the single independent variable is varied in more than two ways (and thus there are more than two values of the independent variable), the more-than-two randomized-groups or the more-than-two-matched-groups design (together called the multi-groups design) is recommended. It is commonly observed that in many behavioural researches, the investigator is faced with a problem in which he is required to manipulate two or more than two independent variables simultaneously. The experimental design suited in such a situation is technically known as the *factorial design*. A factorial design, then, may be defined as a design in which the selected values of two or more independent variables are manipulated in all possible combinations so that their independent as well as interactive effects upon the dependent variable may be studied. The theme of the factorial design is that the different subgroups of subjects work under every possible combination of the factors (or the independent variables) of the design. In brief, a factorial design has the following three main characteristics:

1. Two or more independent variables are manipulated in all possible combinations.
2. For a design to be called factorial, different subgroups or subjects must serve under every possible combination of the independent variables. As far as possible an equal number of subjects in all subgroups is preferred, although this is not an essential condition for a factorial design. An equal number of subjects is preferred because it facilitates statistical computations.
3. The factorial design enables the experimenter to study the independent as well as the interactive effect of the two or more independent variables.

Let us take an example to illustrate the meaning of factorial design. Suppose the experimenter wants to know: Do noise and illumination affect the rate of learning of a list of 15 consonant syllables? Obviously, there are two independent variables: one is noise and the other is illumination. Let us denote the first independent variable as *A* and the second independent variable as *B*. The dependent variable is the rate of learning. Further, suppose that noise is manipulated in two ways: high-noise condition and low-noise condition. Thus there are two levels of *A*. Let the high-noise condition be called A_1 and the low-noise condition A_2. Likewise, the illumination level is manipulated in two ways: the high-illumintion level and the low-illumination level. Thus, there are two levels of *B* also. Let the high-illumination level be called B_1 and the low-illumination level be called B_2. Thus we see that both the independent variables have been manipulated in two ways. The resulting factorial design is 2 × 2. There will be four maximum possible combinations of the experimental conditions generated by these two manipulated variables. These conditions are A_1B_1 (or high-noise as well as high-illumination level condition), A_2B_1 (or low-noise as well as high-illumination level condition), A_1B_2 (high-noise as well as low-illumination level condition) and A_2B_2 (low-noise as well as low-illumination level condition). These four possible combinations have been shown in Table 21.6. When the possible combinations of the independent variables have been prepared, the next step is to randomly assign the subject to different treatment combinations. Let us suppose there are 40 subjects available for the purpose. So, these 40 subjects will be randomly assigned to the four cells of Table 21.6. The random assignment may be done with the help of the table of random numbers. We can, then, start at any row (or column) in the table of random numbers and go on assigning the subjects to the groups according to the sequence of digits in the table. The process would be repeated until each number of the four numbers (I,II,III and IV) has been repeated ten times. Suppose the sequence of digits in the table of random numbers is 26438. It means that subject no. 1 will be assigned to group II; subject no. 2 will be assigned to group IV; and subject no. 3 will be assigned to group III. The digits 6 and 8 will be dropped as we have only four groups. When all the 40 subjects have been randomly assigned to four subgroups, the four treatment combinations may also be randomly assigned so that each group receives one treatment combination. Randomization of subjects into different experimental conditions sometimes becomes a difficult task. This is particularly true when the organismic variables like age, sex, achievement, intelligence, anxiety, etc., are used as the independent variables. Ideally, for complete randomization of subjects in factorial design, independent variables should belong to the category of active variables. In the above-described experiment, the independent variables belong to the category of active variables. In a 2 × 2 factorial design, sometimes it may be found that *A* is an active variable and *B* is the organismic variable or attribute variable. Let us say that *B* stands for the sex variable and *A* stands for noise in any 2 × 2 factorial experiment. The two levels of B are males (B_1) and females (B_2), and the two levels of A are high-noise condition (A_1) and the low-noise condition (A_2). Since *B* is the organismic variable, we cannot assign subjects to B_1 and B_2 at random. However, all male subjects may be randomly assigned to A_1B_1 and A_2B_1 experimental conditions. Likewise, all female subjects may randomly be assigned to A_1B_2 and A_2B_2 experimental conditions. If A, like B, also happens to be an organismic variable, it is risky to use a factorial design.

Table 21.6 A 2 × 2 factorial design for two independent variables

		Noise (*A*)	
		High A_1	Low A_2
Illumination (*B*)	High B_1	A_1B_1	A_2B_1
	Low B_2	A_1B_2	A_2B_2

Types of Factorial Design

A factorial design is directly classified on the basis of the number of independent variables used. However, we shall limit our discussion to factorial design up to three variables. For a discussion of factorial design using more than three variables, the reader is referred to Edwards (1968), Winer (1971), Ostle & Mensing (1975).

Factorial Design with Two Independent Variables

The type of factorial design that we have discussed above is one in which there are two independent variables, each having two levels. Hence, it was referred to as a 2 × 2 factorial design. Likewise, a factorial design with two independent variables may be of a 3 × 2, 3 × 3, 4 × 3, 4 × 4, etc., type. In this way a generalized factorial design for two independent variables may be written as a *K* × *L* factorial design in which *K* stands for the first independent variable and *L* stands for the second independent variable. The value of *K* and *L* indicates the number of ways in which the first and the second independent variables have been manipulated. These ways are known as *levels* of the independent variable. Thus, a 3 × 2 factorial design indicates that the first independent variable has three levels and the second design, there will be six experimental conditions. A 3 × 3 design, similarly, indicates that each of the two independent variables has three levels. Such a design would have nine experimental or treatment combinations.

The simplest factorial design with two independent variables is a 2 × 2 design. Hence, we shall discuss this design in detail by referring to the experiment cited earlier. This experiment had two independent variables and one dependent variable. The two independent variables were noise and illumination, and the dependent variable was the rate of learning. Here the experimenter would have three basic questions before him:

1. Does noise affect the rate of learning?
2. Do the levels of illumination affect the rate of learning?
3. Is there any interaction between the noise and the levels of illumination?

Had the experimenter been interested in only the first two questions, a factorial design (a 2 × 2 design) would not have been necessary. These questions can be answered directly by following a two-randomized-groups design. Since the experimenter is also interested in knowing simultaneously the interaction of the noise and the illumination with respect to the dependent variable scores, a factorial design becomes a necessity.

Before we go to the details of a 2 × 2 factorial design, let us consider the concept of interaction from a more scientific point of view. First, let us illustrate the concept of interaction with an example. Suppose the problem is: Which of the two is more convenient for a person—riding a larger-wheel bicycle or riding a smaller-wheel bicycle? Obviously, the immediate answer will be one of the two: (a) a larger-wheel bicycle, or (b) a smaller-wheel bicycle. However, in addition to these answers, there is the possibility for a third answer, that is, 'it depends upon the body build.' In other words, a tall person may find riding a larger-wheel bicycle more convenient than a smaller-wheel bicycle but a short person may find riding a smaller-wheel bicycle more convenient than a larger-wheel bicycle. It is obvious, thus, that the third answer depends upon the body build of the person who is riding the bicycle. This is the basic notion of the interaction. We can say the act of riding a smaller-wheel bicycle or a larger-wheel bicycle interacts with the body build of the rider. Thus an interaction may be said to exist between the two independent variables when changes produced in the dependent-variable score by one independent variable are also determined by the other independent variable. In the experiment planned above, if interaction exists between the noise (one independent variable) and the level of illumination (another independent variable), it means that the rate of learning under the different conditions of noise would depend upon the levels of illumination found in those conditions or the rate of learning under different levels of illumination would depend upon the different conditions of noise. If there is no interaction between noise and illumination, it means changes in the rate of learning under the two conditions of noise are not dependent upon the two levels of illumination or the changes in the rate of learning under two levels of illumination are not dependent upon the two conditions of noise.

Statistical Analysis in a 2 × 2 Factorial Design

The statistic that is frequently applied to any factorial design is the Analysis of Variance (ANOVA), which was originally developed by Sir R A Fisher. Let us assume that the experimenter got the data (fictitious) shown in Table 21.7 from the learning experiment described above. The table shows the number of trials taken by each subject in learning a list of 15 consonant syllables with the criterion of one perfect recitation. The method of serial learning was followed by the experimenter. The data in Table 21.7 has been rearranged in Table 21.8 so that the means of the four groups may be put in their appropriate cell.

Table 21.7 Fictitious data obtained in an experiment based upon a 2 × 2 factorial design

Group I ($n_1 = 10$)	Group II ($n_2 = 10$)	Group III ($n_3 = 10$)	Group IV ($n_4 = 10$)
High noise and high illumination	Low noise and high illumination	High noise and low illumination	Low noise and low illumination
15	10	16	10
14	12	18	9
20	10	22	8
22	9	25	7
16	8	26	6
18	12	20	10
20	11	20	11
21	10	18	12
18	10	17	10
17	10	16	10
$\Sigma X = 181$	102	198	93
$\Sigma X^2 = 3339$	1054	4034	895
Mean = 18.1	10.2	19.8	9.3

Table 21.8 Means for the four experimental conditions

	Noise (A)		
	High (A_1)	Low (A_2)	Means
High (B_1)	(A_1B_1) $\Sigma X = 181$ Mean = 18.1 $n_1 = 10$	(A_2B_1) $\Sigma X = 102$ Mean = 10.2 $n_2 = 10$	14.15
Low (B_2)	(A_1B_2) $\Sigma X = 198$ Mean = 19.8 $n_3 = 10$	(A_2B_2) $\Sigma X = 93$ Mean = 9.3 $n_4 = 10$	14.55
Means	18.95	9.75	

ΣX = Total score on the dependent variable

Table 21.9 Illustration of the details of the calcuation of ANOVA in a 2 × 2 factorial design (data from Table 21.7)

Step 1: Correction $= \dfrac{(\Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4)^2}{N} = \dfrac{(181 + 102 + 198 + 93)^2}{40}$

$$= \frac{(574)^2}{40} = 8236.9$$

Step 2: Total $SS = (\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2) - C$

$$= (3339 + 1054 + 4034 + 895) - 8236.9$$

$$= 9322 - 8236.9 = 1085.1$$

Step 3: Among $SS = \dfrac{(\Sigma X_1)^2}{n_1} + \dfrac{(\Sigma X_2)^2}{n_2} + \dfrac{(\Sigma X_3)^2}{n_3} + \dfrac{(\Sigma X_4)^2}{n_4} - C$

$$= \frac{(181)^2}{10} + \frac{(102)^2}{10} + \frac{(198)^2}{10} + \frac{(93)^2}{10} - 8236.9$$

$$= \frac{32761 + 10404 + 39204 + 8649}{10} - 8236.9$$

$$= 9101.8 - 8236.9 = 864.9$$

Step 4: Within SS = Total SS – Among $SS = 1085.1 - 864.9 = 220.2$

Step 5: SS between two levels of A (or the first independent variable)

$$\frac{(\Sigma X_1 + \Sigma X_3)^2}{n_1 + n_3} + \frac{(\Sigma X_2 + \Sigma X_4)^2}{n_2 + n_4} - C$$

$$= \frac{(181 + 198)^2}{10 + 10} + \frac{(102 + 93)^2}{10 + 10} - 8236.9$$

$$= \frac{(379)^2}{20} + \frac{(195)^2}{20} - 8236.9$$

$$= 9083.3 - 8236.9 = 846.4$$

Step 6: SS between the two levels of B (or the second independent variable)

$$= \frac{(\Sigma X_1 + \Sigma X_2)^2}{n_1 + n_2} + \frac{(\Sigma X_3 + \Sigma X_4)^2}{n_3 + n_4} - C = \frac{(181 + 102)^2}{10 + 10} + \frac{(198 + 93)^2}{10 + 10} - 8236.9$$

$$= \frac{(283)^2}{20} + \frac{(291)^2}{20} - 8236.9 = 8238.5 - 8236.9 = 1.6$$

Step 7: Interaction SS = Among SS – between SS for A – between SS for B

$$= 864.9 - 846.4 - 1.6 = 864.9 - 848 = 16.9$$

The initial steps involved in calculation of the analysis of variance of a factorial design are the same as those found for a randomized-groups design (see Chapter 23). The first step is to compute the total sum of squares (SS) and subsequently, divide it among SS (see Step 3 of Table 21.9) and the within SS (see Step 4 of Table 21.9). Prior to computing the total sum of squares, we need to compute the correction value (see Step 1). The among SS gives an indication regarding whether or not the groups differ. If one is simply interested in knowing whether or not the different groups under study differ significantly, one need not undertake a factorial design. An overall F test will not answer this question. But as has been stated earlier, in the above experiment there were specific purposes. The experimenter was interested in whether or not variation in each independent variable affected the dependent-variable score and whether there was any significant interaction. Hence, it is essential to divide the among SS into three components: between SS for A, between SS for B, and the $A \times B$ interaction. In order to know whether or not there is significant difference between high-noise condition (A_1) and low-noise condition (A_2), the SS between these two values was computed (see Step 5). Likewise, for knowing whether or not there is a significant difference between the high level of illumination (B_1) and the low level of illumination (B_2), the SS between these two values was computed (see Step 6). The $A \times B$ interaction was easily calculated by subtracting the values of between SS for A and the between SS for B from the among SS (see Step 7). The calculation of the interaction SS completes the computation of the different sums of squares. It must be noted that all the numerical values from step 1 through 7 must be positive. In case a negative value comes at any of these steps, the experimenter should correct the error before he proceeds ahead. After the computation of the different sums of squares is over, the next step is to prepare a complete summary for the analysis of variance as given in Table 21.10.

Table 21.10 A summary of the complete ANOVA in a 2 × 2 factorial design

Source of variation	Sum of squares	DF	Mean square or variance	F
Among all groups	864.9	3	288.3	47.11*
Within groups	220.2	36	6.12	
		...		
Between noise (A_1, A_2)	846.4	1	846.4	138.30
Between illumination (B_1, B_2)	1.6	1	1.6	0.26
Interaction ($A \times B$)	16.9	1	16.9	2.76
Within treatments (error)	220.2	36	6.12	
Total	1085.1	39		

* Not essential for factorial design.

The entries above the dotted line in Table 21.10 may be omitted in a factorial analysis of variance. However, if the purpose of the experimenter is to know the overall differences among all the groups, they should not be avoided. The sums of squares for all the six rows of Table 21.10 have already been calculated in Table 21.9. The formulas for calculating *dfs* are as follows:

Source	*df*	
Among (or between) all groups	$K-1$	(21.1)
Within groups (or within treatments)	$N-K$	(21.2)
Interaction	*df* for $A \times$ *df* for B	(21.3)
	Total $df = N-1$	

In eqs. 21.1 to 21.3, *K* refers to the number of groups and *N* refers to the total number of subjects participating in the experiment. In our example, $N = 40$ and $K = 4$. That is why *df* for among *SS* is $K-1=4-1=3$ (this *df* is based upon all the four groups); and *df* for within *SS* is $N-K=40-4=36$. As among *SS* is divided into three parts, namely, between *SS* for *A* (first independent variable), between SS for *B* (second independent variable) and interaction SS, so also the *df* for among *SS* is divided into corresponding three parts. There are two groups for *A*. Hence, *df* for *A* is $K-1=2-1=1$. Similarly, there are two groups for *B* and accordingly, *df* for *B*, by the same formula, is 1. The *A* effect and *B* effect are called the main effects. Now for the interaction *df* which, in a 2 × 2 factorial design is equal to *df* for *A* ×*df* for *B*. Hence, $1 \times 1 = 1$. This is an example of a two-variable interaction which is called a first-order interaction. Similarly, a three-variable interaction ($A \times B \times C$) and a four-variable interaction ($A \times B \times C \times D$) are known as second-order interaction and third-order interaction respectively. As a check of calculation, when the *df*s for A effect, *B* effect, interaction effect and within treatments are added together they should yield a value equal to $N-1$, which becomes the total *df*. In the randomized-groups design the *df* for the among *SS* plus the *df* for within treatments should be equal to $N-1$, which is again a total *df*. The mean square or variance in each of the six rows of Table 21.10 has been obtained by dividing the respective sum of squares by the respective *df*. For example, within treatments or groups, the row has 220.2 as sum of squares and 36 as *df*. When the former is divided by the latter, we get 6.12 as a mean square. Assuming a fixed model, the values of four *F* ratios have been calculated by using the within variance as a denominator in all cases. For example:

$$F = \frac{288.3}{6.12} = 47.11$$

$$F = \frac{864.4}{6.12} = 138.30$$

$$F = \frac{1.6}{6.12} = 0.26$$

$$F = \frac{16.9}{6.12} = 2.76$$

Interpretation of *F* ratios: There were three specific null hypotheses in the 2 × 2 factorial design described above.

1. There is no difference between the means of the two conditions of noise. In other words, the means of A_1 and A_2 do not differ.
2. There is no difference between the means of the two conditions of illumination. In other words, the means of B_1 and B_2 do not differ.
3. There is no interaction between the two independent variables, that is, between noise and illumination.

These three null hypotheses were tested by the last three *F*s. The first *F* is intended to test a general null hypothesis which indicated no difference between all the four groups under study. As this can be avoided in a factorial design, let us drop its discussion. The interested reader can study this *F* in the same way as for the *F* test in a randomized-groups design (see Chapter 23).

Let us turn to the first null hypothesis which relates to the effect of noise upon the rate of learning. Let the alpha level for each *F* test be 0.05. For the *F* relating to *A* effect, we have 1 *df* for numerator and 36 *df* for denominator. The value of *F* required at 0.5 level of significance with 1 df_1 and 36 df_2 is 4.11 (see Edwards 1968*). Since the obtained *F* of 138.30 exceeds this value, we can reject the null hypothesis and conclude that two conditions of noise, averaged over the two levels of *B*, produced a significant difference in the rate of learning. As the mean for high-noise condition (18.95) is larger than the mean for low-noise condition (9.75), the experimenter may conclude that high noise retards the rate of learning because the subjects took more trials in reaching the criterion. The second null hypothesis relates to the effect of illumination upon the rate of learning. Again, we have 1 *df* for the numerator and 36 *df* for the denominator for the obtained *F* of 0.26. As this *F* is less than 1, without consulting the probability table for *F*, it can immediately be taken to be not significant. Hence, the second null hypothesis is accepted. We conclude, therefore, that two conditions of illumination, averaged over the two levels of *A*, are not producing the significant difference in the rate of learning. Thus the obtained mean difference of 0.40 between B_1 and B_2 is due to the chance factor.

After the interpretation of the main effects is over, the interpretation of interaction effects is taken up. In a 2 × 2 factorial design, there is only a first-order interaction. The interaction *F* is 2.76 and we have 1 *df* for the numerator and 36 *df* for the denominator for this *F*. As this value of *F* is less than the required value of *F* (that is, 4.11) to be significant at 0.05 level, we accept our third null hypothesis, too. As the null hypotheses for *A* × *B* interaction is accepted, we may say that the *A* effect, that is, the difference between A_1 and A_2 is not dependent upon *B*. In other words, we have approximately the same difference between A_1 and A_2, irrespective of the levels of *B*. Thus it is finally concluded that the rate of learning in high-noise condition and low-noise condition is not affected by the given level of illumination. When nonsignificant interaction is graphically shown, it presents the parallel lines. This can be easily checked by Figure 21.4. When the interaction is significant, it yields nonparallel lines in a graphical representation.

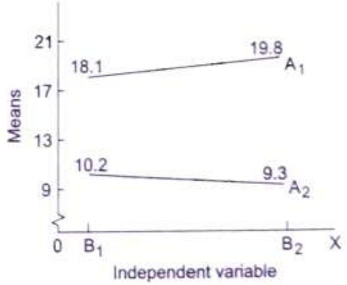

Fig. 21.4 Means for levels of *A* at each level of *B*

The factorial design discussed above deals with only two variables. It can, however, be extended to three or any number of independent variables. Due to the reason of complexity involved in calculation, we shall limit our discussion for three independent variables only. The simplest such factorial design in which we have a total of eight treatment combinations is a 2 × 2 × 2 factorial design. In this design, we have three factors (or independent variables) with two levels for each factor. The general case for a factorial design utilizing three independent variables is a $K \times L \times M$ design where K, L and M stand for the first, second and third independent variables respectively.

---

* A L Edward's *Experimental Design in Psychological Research*, Holt, Rinehart & Wilson, 1968, Table VIII, p. 422

Let us illustrate a $2 \times 2 \times 2$ factorial experiment with an example from a randomized-groups design. Suppose the dependent variable in an experiment is the measure of retention of a learnt task. The experimenter wants to investigate the influence of three independent variables and their interactions, if any, upon the retention score. Let the first independent variable be the mode of presentation of the materials. This may be called A and this independent variable is manipulated in two ways—visual and auditory. That is, in one case the entire list is presented to the subject and they themselves read it. This constitutes the visual mode of presentation. In another case, the entire list is read aloud to them and this constitutes the auditory presentation of the materials. Let the visual mode be called A_1 and the auditory mode be called A_2. The second independent variable is the number of presentations and this factor may be designated as B. This factor is also varied by the experimenter in two ways: one complete presentation of the list and two complete presentations of the list. Let one presentation be designated as B_1 and the two presentations be B_2. The third factor is the condition in which materials are presented. Let this condition be called C and be manipulated in two ways: noisy condition and quiet condition. We designate noisy condition as C_1 and quiet condition as C_2. Thus, all three factors, that is, A, B and C are manipulated in two ways and therefore, each factor has two levels. There will be eight treatment combinations and a treatment will be obtained by selecting one level from each independent variable. For example, $A_1 B_2 C_1$ is one treatment and consists of a visual mode having two presentations in the noisy condition. Similarly, the other seven treatment combinations can be interpreted. Suppose, further, that 80 randomly selected subjects from a population of students belonging to an intermediate class have been made available to the experimenter. Assuming a randomized-groups design, these 80 subjects are randomly assigned to eight treatment combinations or groups so that each group has 10 subjects. The dependent variable score (fictitious) as obtained in this experiment was presented in Table 21.11. The data of Table 21.11, may be reduced like those shown in Table 21.12 so that the means of eight groups may be shown in the appropriate cell, which in turn, provides facilitation in further calculation. The reduced data are shown in Table 21.12.

Now let us return to the calculation of ANOVA in a $2 \times 2 \times 2$ factorial experiment.

As stated earlier, the initial steps of calculating ANOVA are similar to those found in the 2×2 factorial design presented in Table 21.9.

Table 21.11 Fictitious dependent variable scores in a $2 \times 2 \times 2$ factorial experiment

Gr. I $A_1B_1C_1$	Gr. II $A_1B_1C_2$	Gr. III $A_2B_1C_1$	Gr. IV $A_2B_1C_2$	Gr. V $A_1B_2C_1$	Gr. VI $A_1B_2C_2$	Gr. VII $A_2B_2C_1$	Gr. VIII $A_2B_2C_2$
10	10	3	4	16	17	12	15
8	8	2	6	16	15	13	14
7	7	6	8	14	20	13	16
7	6	6	7	15	18	10	16
6	10	3	7	15	18	10	17
10	10	3	8	16	19	10	15
10	10	5	8	16	14	12	15
9	12	6	6	15	18	12	16
9	11	6	8	14	17	11	17
8	9	5	8	14	17	10	16
$\Sigma X = 84$	93	45	70	151	173	113	157
$\Sigma X^2 = 724$	895	225	506	2287	3021	1291	2473
Mean = 8.4	9.3	4.5	7.0	15.1	17.3	11.3	15.7

---

* A factorial experiment can be conducted not only with the randomized-groups design but also with randomized-block design and Latin-square design.

Table 21.12 Means for the eight experimental conditions in a $2 \times 2 \times 2$ factorial design

	A_1		A_2	
	C_1	C_2	C_1	C_2
B_1	$(A_1B_1C_1)$ $\Sigma X = 84$ Mean = 8.4 $n_1 = 10$	$(A_1B_1C_2)$ $\Sigma X = 93$ Mean = 9.3 $n_2 = 10$	$(A_2B_1C_1)$ $\Sigma X = 45$ Mean = 4.5 $n_3 = 10$	$(A_2B_1C_2)$ $\Sigma X = 70$ Mean = 7.0 $n_4 = 10$
B_2	$(A_1B_2C_1)$ $\Sigma X = 151$ Mean = 15.1 $n_5 = 10$	$(A_1B_2C_2)$ $\Sigma X = 173$ Mean = 17.3 $n_6 = 10$	$(A_2B_2C_1)$ $\Sigma X = 113$ Mean = 11.3 $n_7 = 10$	$(A_2B_2C_2)$ $\Sigma X = 157$ Mean = 15.7 $n_8 = 10$

ΣX = Total score

Table 21.13 Illustration of the details of the calculation of ANOVA in a $2 \times 2 \times 2$ factorial design (data are from Table 21.11 and Table 21.12)

Step 1: Correction

$$\text{Correction} = \frac{(\Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4 + \Sigma X_5 + \Sigma X_6 + \Sigma X_7 + \Sigma X_8)^2}{N}$$

$$= \frac{(84 + 93 + 45 + 70 + 151 + 173 + 113 + 157)^2}{80} = \frac{(886)^2}{80} = 9812.45$$

Step 2: Total *SS*

$$\text{Total } SS = (\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 + \Sigma X_5^2 + \Sigma X_6^2 + \Sigma X_7^2 \,|\, \Sigma X_8^2) - C$$

$$= (724 + 895 + 225 + 506 + 2287 + 3021 + 1291 + 2473) - 9812.45$$

$$= 11422 - 9812.45 = 1609.55$$

Step 3: Among *SS*

$$\text{Among } SS = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + \frac{(\Sigma X_4)^2}{n_4} + \frac{(\Sigma X_5)^2}{n_5} + \frac{(\Sigma X_6)^2}{n_6} + \frac{(\Sigma X_7)^2}{n_7} + \frac{(\Sigma X_8)^2}{n_8} - C$$

$$= \frac{(84)^2}{10} + \frac{(93)^2}{10} + \frac{(45)^2}{10} + \frac{(70)^2}{10} + \frac{(151)^2}{10} + \frac{(173)^2}{10} + \frac{(113)^2}{10} + \frac{(157)^2}{10} - 9812.45$$

$$= \frac{7056 + 8649 + 2025 + 4900 + 22801 + 29929 + 12769 + 24649}{10} - 9812.45$$

$$=\frac{112778}{10}-9812.45=1465.35$$

Step 4: Within *SS* = Total *SS* – Among *SS* = 1609.55 – 1465.35 = 144.20

Step 5: *SS* between two levels of A (or first independent variable):

$$=\frac{(\Sigma X_1+\Sigma X_2+\Sigma X_5+\Sigma X_6)^2}{n_1+n_2+n_5+n_6}+\frac{(\Sigma X_3+\Sigma X_4+\Sigma X_7+\Sigma X)_8^2}{n_3+n_4+n_7+n_8}-C$$

$$=\frac{(84+93+151+173)^2}{10+10+10+10}+\frac{(45+70+113+157)^2}{10+10+10+10}-9812.45$$

$$=\frac{(501)^2}{40}+\frac{(385)^2}{40}=9812.45$$

$$=6275.025+3705.625-9812.45=168.200$$

Step 6: *SS* between two levels of *B* (second independent variable):

$$=\frac{(\Sigma X_1+\Sigma X_2+\Sigma X_3+\Sigma X_4)^2}{n_1+n_2+n_3+n_4}+\frac{(\Sigma X_5+\Sigma X_6+\Sigma X_7+\Sigma X_8)^2}{n_5+n_6+n_7+n_8}-C$$

$$=\frac{(84+93+45+70)^2}{10+10+10+10}+\frac{(151+173+113+157)^2}{10+10+10+10}-9812.45$$

$$=\frac{(292)^2}{40}+\frac{(594)^2}{40}-9812.45$$

$$=2131.6+8820.9-9812.45=1140.05$$

Step 7: *SS* between two levels of *C* (third independent variable):

$$=\frac{(\Sigma X_1+\Sigma X_3+\Sigma X_5+\Sigma X_7)^2}{n_1+n_3+n_5+n_7}+\frac{(\Sigma X_2+\Sigma X_4+\Sigma X_6+\Sigma X_8)^2}{n_2+n_4+n_6+n_8}-C$$

$$=\frac{(84+45+151+113)^2}{10+10+10+10}+\frac{(93+70+173+157)^2}{10+10+10+10}-9812.45$$

$$=\frac{(393)^2}{40}+\frac{(493)^2}{40}-9812.45=3861.225+6076.225-9812.45=125$$

Step 8: Interaction *SS*: *A* × *B*

$$=\frac{[(a+d)-(b+c)]^2}{4n}=\frac{[(177+270)-(115+324)]^2}{(4)(20)}=0.8$$

Step 9: Interaction *SS*: *A* × *C*

$$=\frac{[(a+d)-(b+c)]^2}{4n}=\frac{[(235+227)-(158+266)]^2}{(4)(20)}=18.05$$

Step 10: Interaction *SS*: *B* × *C*

$$=\frac{[(a+d)-(b+c)]^2}{4n}=\frac{[(129+330)-(264+163)]^2}{(4)(20)}=12.8$$

Step 11: Interaction *SS*: *A* × *B* × *C*

Among *SS* – (*SS* for *A* + *SS* for *B* + *SS* for *C* + Interaction *SS* for *A* × *B* + Interaction SS for *A* × *C* + Interaction SS for *B* × *C*)

$$=1465.35-(168.2+1140.05+125+0.8+18.05+12.8)$$

$$=1465.35-1464.90=0.45$$

An inspection of Table 21.13 reveals that the initial steps (that is, up to Step 4) for calculation of analysis of variance is similar to those found in the randomized-groups design. The among *SS* tells the overall differences between the groups under study. In a 2 × 2 × 2 factorial design the among *SS* (or treatment sum of squares) is divided into several parts: between *SS* for *A*, between *SS* for *B*, between *SS* for *C*, interaction *SS* for *A* × *B*, interaction *SS* for *A* × *C*, interaction *SS* for *B* × *C* and interaction *SS* for *A* × *B* × *C*. The between *SS* for *A* is computed to know whether or not there is significant difference between the visual mode (A_1) and auditory mode (A_2) of presentation. Likewise, between *SS* for *B* is computed to know whether or not there is significant difference between one presentation (B_1) or two presentations (B_2) of the list, so also the between *SS* for *C* is computed to know whether or not there is difference between noisy condition (C_1) and quiet condition (C_2).

The interaction *SS* was computed with the help of the following formula:

$$\frac{[(a+d)-(b+c)]^2}{4n} \qquad (21.4)$$

Equation 21.4 is applicable when the data of the factorial experiment has been reduced to a 2 × 2 table and *a*, *b*, *c* and *d* are the four cells of a 2 × 2 table. A schematic representation of a 2 × 2 table for computing the interaction SS with the help of Eq. 21.4 is presented below.

Table 21.14 A schematic representation of a 2 × 2 table for computation of the interaction SS

	A_1	A_2
B_1	*a*	*b*
B_2	*c*	*d*

Thus, for computing three first-order interactions the data (Σ*X*) of Table 21.12 has been reduced as shown in Table 21.15.

Each cell in Table 21.15 represents the sum of 20 scores or observations.

Table 21.15 A 2 × 2 table for A × B, A × C and B × C interactions

(i) 2 × 2 table for *A* × *B* interaction:

	A_1	A_2
B_1	84 + 93 = 177	70 + 45 = 115
B_2	151 + 173 = 324	113 + 157 = 270

(ii) 2 × 2 table for $A \times C$ interaction:

	A_1	A_2
C_1	$84 + 151 = 235$	$45 + 113 = 158$
C_2	$93 + 173 = 266$	$70 + 157 = 227$

(iii) 2 × 2 table for $B \times C$ interaction:

	B_1	B_2
C_1	$84 + 45 = 129$	$151 + 113 = 264$
C_2	$93 + 70 = 163$	$173 + 157 = 330$

Based upon these scores, the interaction SS has been calculated (see steps 8, 9 and 10 of Table 21.13). When the sum of squares for $A, B, C, A \times B, A \times C$ and $B \times C$ has been calculated, the second-order interaction ($A \times B \times C$) can be easily calculated by adding the above six sums of squares together and subtracting the summed value from the among SS. The $A \times B \times C$ interaction can, however, be calculated directly but owing to its lengthy process of calculation the author does not think proper to elaborate upon it here. Once more it must be added that all values from steps 1 through 11 must be positive. If any one of them happens to be negative, the investigator should try to find out the error before he proceeds further. After calculating $A \times B \times C$ interaction, the summary of the complete analysis of variance is set up in a manner presented in Table 21.16. The entries above the dotted line in Table 21.16 may be omitted if the purpose of the

Table 21.16 A summary of the complete ANOVA in a 2 × 2 × 2 factorial experiment

Source	Sum of squares	df	Mean square	F
Among all groups	1465.35	7	209.336	104.511
Within groups	144.20	72	2.003	
A: Mode	168.20	1	168.20	83.974
B: Number	1140.05	1	1140.05	569.171
C: Condition	125.00	1	125.00	62.406
$A \times B$: Mode × Presentation	0.80	1	0.80	0.399
$A \times C$: Mode × Condition	18.05	1	18.05	9.011
$B \times C$: Presentation × Condition	12.80	1	12.80	6.390
$A \times B \times C$: Mode × Presentation × Condition	0.45	1	0.45	0.224
Error: Within groups	144.20	72	2.003	
Total:	1609.55	79		

experimenter is not to determine the overall differences among the eight groups. This is particularly true in the case of a factorial experiment. In a factorial experiment having three independent variables, the among SS is divided into seven parts. Accordingly, dfs can also be divided into seven parts. The df for among groups and within groups is calculated by eqs. 21.1 and 21.2 respectively. For calculation of the sum of squares for A, the eight groups are reduced to only two groups, each consisting of 40 subjects. Thus B and C independent variables are temporarily ignored. As two groups form for A, the df for the A effect by Eq. 21.1 is 2 − 1 = 1. Similarly, for the sum of squares for B, the two other independent variables (that is A, and C) are temporarily ignored and for the sum of squares for C, two independent variables that is, A and B are temporarily ignored. In each case, all the groups are reduced to two. In B effect, one group consisting of 40 observations is for B_1 and another group consisting of another 40 observations is for B_2. Similarly, in the C effect, one group consisting of 40 observations is for C_1 and another group consisting of another 40 observations is for C_2. The df for B effect, by Eq. 21.1, is 2 − 1 = 1 and df for C effect by the same formula is 2 − 1 = 1. For calculating the df for $A \times B \times C$ interaction, we multiply the df for A with the df for B with the df for C. As we have 1 df for each of three main effects, df for $A \times B \times C$ interaction is also 1. Similarly, the dfs for three first-order interactions have been calculated. The mean square or variance for each source is calculated by dividing the sums of square with their corresponding df. For example, the within SS is 144.20 which, if divided by 72, yields 2.003 as the mean square. The df for within groups has been calculated by Equation 21.2. As the above example illustrates fixed model of analysis of variance, we can resasonably choose the within variance to act as the error term in calculating the F ratios. Therefore, the F ratios for main effects and interaction effects have been calculated as follows:

$$F \text{ for } A = \frac{168.20}{2.003} = 83.974$$

$$F \text{ for } B = \frac{1140.05}{2.003} = 569.171$$

$$F \text{ for } C = \frac{125}{2.003} = 62.406$$

$$F \text{ for } A \times B = \frac{0.80}{2.003} = 0.399$$

$$F \text{ for } A \times C = \frac{18.05}{2.003} = 9.011$$

$$F \text{ for } B \times C = \frac{12.80}{2.003} = 6.390$$

$$F \text{ for } A \times B \times C = \frac{0.45}{2.003} = 0.224$$

Interpretation of F Ratios

In the above factorial experiment the alternative hypothesis* (AH) and null hypothesis (NH) were as follows:

(i) There will be significant difference between the means of the visual mode (A_1) and auditory mode (A_2)—AH. There will be no significant difference between the means of visual mode (A_1) and auditory mode (A_2)—NH.

---

* All alternative hypotheses have been developed in the two-tailed manner. One-tailed-hypotheses may also be developed.

(ii) There will be significant difference between the means of one presentation (B_1) and two presentations (B_2)—AH.

There will not be a significant difference between the means of one presentation and two presentations—NH.

(iii) There will be significant difference between the means of noisy condition (C_1) and quiet condition (C_2)—AH.

There will not be significant difference between the means of noisy condition and quiet condition—NH.

(iv) There will be interaction between the two independent variables, namely, the methods of presentation (*A*) and the number of presentations (*B*)—AH.

There will be no interaction between the mode and number of presentations—NH.

(v) There will be interaction between the modes of presentation (*A*) and conditions (*C*) in which materials are presented—AH.

There will be no interaction between the modes of presentation and conditions (*C*) in which materials are presented—NH.

(vi) There will be interaction between the number of presentations (*B*) and conditions (*C*) in which materials are presented—AH.

There will be no interaction between number of presentations and conditions in which materials are presented—NH.

(vii) There will be interaction among the three independent variables, namely, mode (*A*), number (*B*), and condition (*C*)—AH.

There will not be any significant difference among the interaction of all the three independent variables, namely, mode (*A*), number (*B*) and condition (*C*)—NH.

The first three hypotheses are related to the main effects, the last hypothesis is concerned with the second-order interaction and the remaining three hypotheses are concerned with the first-order interaction. Let the alpha level be set at 0.05 level. The *F* for the first main effect, that is, for *A* effect is 83.974 for which we have 1 *df* in the numerator and 72 *df* in the denominator. The required *F* value for $df = 1/72$ at 0.05 level = 3.975 (by interpolation). As the obtained *F* exceeds this value, we reject the null hypothesis and accept the alternative hypothesis. This *A* effect represents the comparison between the means for the visual mode and the auditory mode, averaged over the two levels of *B* and the two levels of *C*. The mean for the visual mode is $501/40 = 12.525$ and the mean for the auditory mode is $385/40 = 9.625$. This obviously means that the visual mode of presentation results in a more superior average retention than the auditory mode of presentation.

Similarly, *B* effect corresponds to the comparison between the means of one presentation and two presentations *averaged over* the two levels of *A* and the two levels of *C*. The obtained *F* for *B* effect is 569.171 and we have 1 *df* in the numerator and 72 *df* in the denominator. As the obtained *F* exceeds the required value of *F* (3.975) at 0.05 level, the null hypothesis is rejected and the alternative hypothesis is accepted. Thus, there is a significant difference between the means of one presentation and two presentations. As the mean for one presentation ($292/40 = 7.3$) is far less than the mean for two presentations ($594/40 = 14.85$), we conclude that two presentations of the task definitely result in a superior average retention than one presentation.

The *C* effect represents comparison between the means of the noisy condition and the quiet condition averaged over the two levels of *A* and the two levels of *B*. The obtained *F* for *C* is 62.406, which has 1 *df* in the numerator and 72 *df* in the denominator. The obtained *F* exceeds the required *F* for the above *df*s at 0.05 level. Hence the third null hypothesis is also rejected. Again, the alternative hypothesis is accepted. The mean for the noisy condition is $393/40 = 9.825$ and the mean for the quiet condition is $493/40 = 12.325$. As the mean difference is significant, we conclude that the quiet condition definitely results in a superior average retention than the noisy condition.

Of the three first-order interactions, one is not significant and two are significant. The $A \times B$ interaction is not significant because the obtained *F* for this interaction is less than 1. The nonsignificant nature of the $A \times B$ interaction is also checked from the two parallel lines in Figure 21.5. As *F* for $A \times B$ interaction is not significant, the null hypothesis, rather than the alternative hypothesis, is accepted. The nonsignificant $A \times B$ interaction suggests that the difference between the means of visual mode (A_1) and the auditory mode (A_2) is not dependent upon either one presentation (B_1) or two presentations (B_2) of the task. Or, it may be interpreted in reversed direction. But $A \times C$ interaction is significant because the obtained *F* for this interaction is 9.011, which has 1 *df* in the numerator and 72 *df* in the denominator. As this obtained *F* exceeds the required *F* at 0.05 level for the above *df*s, we reject the null hypothesis and accept the alternative hypothesis. The significant nature of $A \times C$ interaction is also obvious from the two nonparallel lines in Figure 21.6. The significant $A \times C$ interaction suggests that the mean difference between visual mode (A_1) and auditory mode (A_2) is dependent upon (or determined by) the noisy and the quiet condition in which the task is presented. Thus *A* effect is not the same for the two levels of C. In noisy condition (at C_1) the visual mode of presentation (A_1) yields the mean of $235/20 = 11.75$ and the auditory mode (A_2) yields the mean of $158/20 = 7.9$. The difference between A_1 and A_2 for C_1 is $11.75 - 7.9 = 3.85$. In quiet condition (C_2) the visual mode

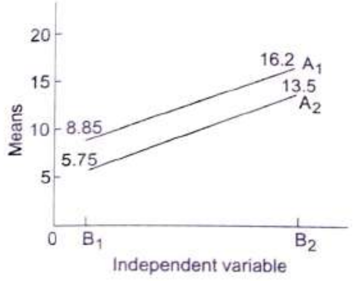

Fig. 21.5 Means for levels of *A* at each level of *B*

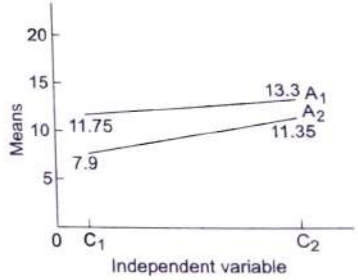

Fig. 21.6 Means for levels of *A* at each level of *C*

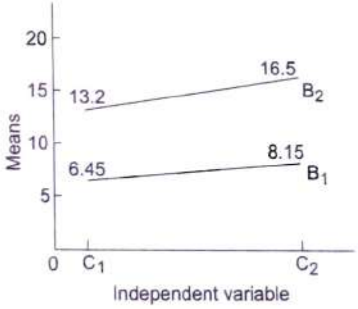

Fig. 21.7 Means for levels of *B* at each level of *C*

yielded a mean of 266/20 = 13.3 and the auditory mode yielded a mean of 277/20 = 11.35. The mean difference between A_1 and A_2 is 13.3 − 11.35 = 1.95. A simple comparison of the mean difference between A_1 and A_1 for C_1 as well as for C_2 reveals that the magnitude of difference is not the same for these two levels of C, a fact which has made the interaction significant. Like $A \times C$ interaction, $B \times C$ interaction is also significant because the obtained F (6.390) exceeds the required value of F (3.975). This F also has 1 *df* in the numerator and 72 *df* in the denominator. We accept the alternative hypothesis and reject the null hypothesis. The significant nature of $B \times C$ interaction can also be read from two nonparallel lines shown in Figure 21.7. The significant $B \times C$ interaction may be taken to mean that the mean difference between B_1 and B_2 at both levels of C is not the same. In noisy condition (C_1) the mean of one presentation (B_1) is 129/20 = 6.45 and the mean of two presentations (B_2) is 264 / 20 = 13.2. Therefore, we get a mean difference of 6.75 between B_1 and B_2 for C_1. In the quiet condition (C_2) the mean of one presentation (B_1) is 163 / 20 = 8.15 and the mean of two presentations (B_2) is 330 / 20 = 16.50. Thus for C_2 the mean difference between B_1 and B_2 is 8.35. It is obvious that the mean difference between B_1 and B_2 for C_1 is not the same as the mean difference between B_1 and B_2 for C_2. The greater this difference, the greater the probability of the interaction being significant. The $A \times B \times C$ interaction is not significant and hence, our last null hypotheis is not rejected meaning thereby that our last alternative hypothesis is rejected. This second-order interaction or three-factor interaction needs a little more explanation.

Just as the two-factor interaction, for example $A \times B$ interaction, is the symmetrical property of A and B, so also $A \times B \times C$ interaction is regarded as the symmetrical property of A, B and C factors. This means that we can explain $A \times B \times C$ interaction by considering interaction $A \times B$ separately for each level of C or by considering $A \times C$ interaction separately for each level of B or by considering $B \times C$ interaction separately for each level of A. Let us explain $A \times B \times C$ interaction by considering $A \times B$ interaction separately for each level of C. As the $A \times B \times C$ interaction is not significant because the obtained F for this interaction is less than 1, $A \times B$ interaction for C_1 and C_2 is the same. When graphically shown, the figure for $A \times B$ interaction for C_1 and for C_2 should be of the same form. We may also explain the nature of $A \times B \times C$ interaction by grouping $A \times C$ interaction separately for each level of B, or by grouping $B \times C$ interaction separately for each level of A. As we have explained $A \times B$ interaction by considering $A \times B$ interaction for each level of C, the meaning of the second-order interaction is that $A \times B$ interaction is the same for C_1 and C_2. The $A \times B$ interaction for C_1 has been shown in Figure 21.8 and for C_2 in Figure 21.9. For graphically showing the $A \times B \times C$ interaction, the data have been reduced to a 2 × 2 table form as presented in Table 21.17. It is obvious that the forms of the graph in Figures 21.8 and 21.9 are nearly similar, which confirms the fact that $A \times B \times C$ interaction is not significant.

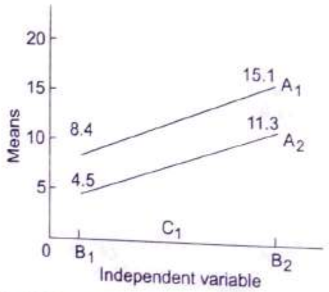

Fig. 21.8 Means for levels of A at each level of B for C_1

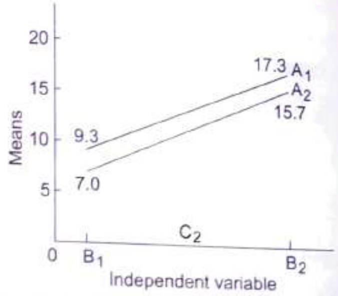

Fig. 21.9 Means for levels of A at each level of B for C_2

Table 21.17 A 2 × 2 table for means of *A* and *B* for each level of *C*

	For C_1			For C_2	
	A_1	A_2		A_1	A_2
B_1	8.4	4.5	B_1	9.3	7.0
B_2	15.1	11.3	B_2	17.3	15.7

Advantages And Disadvantages of Factorial Experiments

The factorial experiments have several advantages over a single IV experimment.

1. In a factorial design two or more than two independent variables are simultaneously manipulated whereas in a single IV experiment, as its name implies, a separate experiment is designed to study the effect of independent variables. Thus a factorial experiment provides economy of time, labour and money.

2. The factorial experiments also permit the evalution of interaction upon the dependent variable. In a single IV experiment, one cannot evaluate the effect of the interacion of the independent variables because only one independent variable is manipulated at a time.

3. The experimental results of a factorial experiment are more comprehensive and can be generalized to a wider range due to the manipulation of several independent variables in one experiment. From this point of view, the single IV experiments suffer a major setback.

4. In factorial experiments, there is an additional gain occurring due to the hidden replication arising from the factorial arrangement itself.

Despite these advantages, a factorial experiment has some disadvantages.

1. Sometimes the experimental set-up and the resulting statistical analyses become so complex that the experimenter may wish to drop this design and return to a single IV experiment. This is especially true when more than three independent variables, each with three or more levels, are to be manipulated together.

2. In factorial experiments when the number of treatment combinations or treatments becomes large, it becomes difficult for the experimenter to select a homogeneous experimental unit (or subject).

3. Sometimes, it happens that some treatment combinations arising out of the simultaneous manipulation of several independent variables become meaningless. Then, the resources spent in those combinations are simply wasted.

Selecting a Correct Error Term in a Factorial F Test

To select a correct error term in ANOVA is an important problem. In the randomized-groups design, the within-groups mean square is used as the correct error term in an *F* test. This is also the usual case in the factorial experiment. But sometimes we do not use within-groups mean square as the error term in a factorial experiment. In order to understand this situation, we shall examine three common types of factorial experiments. The first type is known as a *fixed model*, the second type is known as a *random model* and the third type is known as a *mixed model*. A model, here, refers to the characteristics of the independent variables used in the factorial experiment. Let us discuss these three models separately and illustrate each with the simplest factorial design, that is with a 2 × 2 factorial design.

Fixed Model: In a fixed model of ANOVA, the experimenter has some obvious and particular reasons to select the values of the independent variables. In other words, the experimenter here does not choose the values of the independent variable at random; rather he chooses them with some particular reasons. For example, the experimenter may choose the different levels of noise, light, practice, and the methods of teaching, etc., as the different values of the independent variables for his experiment. He may choose to compare a 10-trial performance with a 100-trial performance; he may choose methods *A* and *B* and compare their impact upon the dependent variable measure. All these illustrate the case of a fixed model. In psychological research, the case of the fixed model is highly common. For the case of a fixed model, the within-groups mean square (*MS*) is always used as the correct error term for all *F* tests. Thus , for a 2 × 2 factorial design we have three *F*s which are calculated as follows:

$$F \text{ for } A \text{ (or first independent variable)} = \frac{MS \text{ for } A}{MS \text{ for } within\text{-}groups}$$

$$F \text{ for } B \text{ (or the second independent variable)} = \frac{MS \text{ for } B}{MS \text{ for } within\text{-}groups}$$

$$F \text{ for } A \times B = \frac{MS \text{ for } A \times B}{MS \text{ for } within\text{-}groups}$$

In a factorial experiment with two independent variables the total sum of squares is divided into four components, namely, the sum of squares for factor *A*, for factor *B*, for interaction *A* × *B* and for within-groups. Based on these four sums of squares, we calculate four mean squares or variances separately and our null hypotheses are tested in terms of *F* tests, which are expressed in terms of the ratio of these variances. Except the mean square for within-group, all other mean squares may contain more than one component of variance. Table 21.18 presents the expected mean square for a fixed model, random model and a mixed model. Look at the fixed model as we are discussing only this model at present. The expected mean square for the *A* factor is expressed by $\sigma^2_W + 2n\sigma^2_A$. This means that the mean square for *A* contains the mean square for within-groups (σ^2_W) and $2n$ times the variance of the *A* factor (σ^2_A). For the present purpose, we can ignore a constant like $2n$ and concentrate on the mean squares or variances only. Since the expected mean square for *A* contains the within variance as well as variance due to *A*, we divide this expected mean square by the within variance. This is the rationale behind using the within variance as the correct error term in a fixed model. The same reasoning applies for the independent variable *B* and hence, the within variance is used as the error term. The expected mean square for interaction *A* × *B* contains the within variance (σ^2_W) and *n* times the interaction variance (σ^2_{AB}). Dropping the constant *n* as usual, we may say that the expected mean square for interaction *A* × *B* is composed of the within variance as well as the interaction variance. When we divide this expected mean square by the within variance, we get the actual *F* value for the interaction variance.

Table 21.18 Expected mean squares for a 2 × 2 factorial design in different models

Source of variation	Expected Mean squares			
	Fixed model (Model-I)	Random model (Model-II)	Mixed model (Model-III) *A* fixed, *B* random	Mixed model (Model-III) *A* random, *B* fixed
Independent variable *A*	$\sigma^2_W + 2n\sigma^2_A$	$\sigma^2_W + 2n\sigma^2_{AB} + 2n\sigma^2_A$	$\sigma^2_W + 2n\sigma^2_{AB} + 2n\sigma^2_A$	$\sigma^2_W + 2n\sigma^2_A$
Independent variable *B*	$\sigma^2_W + 2n\sigma^2_B$	$\sigma^2_W + 2n\sigma^2_{AB} + 2n\sigma^2_B$	$\sigma^2_W + 2n\sigma^2_W$	$\sigma^2_W + 2n\sigma^2_{AB} + 2n\sigma^2_B$
Interaction *A* × *B*	$\sigma^2_W + 2n\sigma^2_{AB}$	$\sigma^2_W + 2n\sigma^2_{AB}$	$\sigma^2_W + 2n\sigma^2_{AB}$	$\sigma^2_W + 2n\sigma^2_{AB}$
Within-groups	σ^2_W	σ^2_W	σ^2_W	σ^2_W

Random Model: When the values of the independent variables are selected at random, we have the case of a random model. In a 2 × 2 factorial design, suppose the experimenter selects the age of subjects and the method of teaching as two independent variables. Now, he may decide to select two levels of each independent variable. He may place reasonable possible values of age, say, ranging from 6 years to 30 years (25 slips), into a box and draw any two slips (which specify two different age levels) at random. Likewise, suppose for example, there are 6 reasonable possible methods of teaching. These may be numbered from 1 to 6, each on a separate slip, and placed in a box. After reshuffling, the experimenter may select any two slips at random. The number thus drawn will represent the two methods of teaching selected at random. The column for a random model in Table 21.18 indicates that the expected mean square for the independent variable $A = \sigma^2_W + n\sigma^2_{AB} + 2n\sigma^2_A$. Dropping the constants $2n$ and n for the present purpose and concentrating only upon the variances, we can say that the expected mean square for *A* consists of the within variance (σ^2_W), the interaction variance (σ^2_{AB}) and the variance due to the factor *A*. Hence, we need here to divide both the interaction variance and the within variance. Since *A* × *B* interaction contains these two variances, that is, the interaction variance and the within variance, we divide the expected mean square for *A* by interaction variance. Similar is the case with the independent variable *B*. Hence, we divide the expected mean square for this variable by the interaction variance. The *A* × *B* interaction contains the within variance (σ^2_W) and the interaction variance (σ^2_{AB}). Hence, we divide it by the within variance and obtain the *F* for the interaction variance. Thus in a random model of a 2 × 2 factorial design, we compute the three *F*s as follows:

$$F \text{ for } A = \frac{MS \text{ for } A}{MS \text{ for interaction}}$$

$$F \text{ for } B = \frac{MS \text{ for } B}{MS \text{ for interaction}}$$

$$F \text{ for } A \times B = \frac{MS \text{ for interaction}}{MS \text{ for within-groups}}$$

Mixed Model: As its name implies, the case for the mixed model is one in which one variable is fixed and another variable is random. Let us first consider a mixed model in which the independent variable *A* is fixed and the independent variable *B* is random. In such a model the expected mean square for the independent variable *A* consists of the within variance (σ^2_W), the interaction variance (σ^2_{AB}) and the variance due to $A(\sigma^2_A)$. As the expected mean square for the *A* × *B* interaction in such a mixed model contains the above first two variances, we divide the expected mean square for the independent variable *A* by the interaction variance and obtain the *F* for *A*. The expected mean square for independent variable *B* contains the within variance $(\sigma^2_W$ and the variance due to $B(\sigma^2_B)$. Hence, we divide this mean square by the within variance and obtain the *F* for *B*. Likewise, the expected mean square for *A* × *B* interaction contains the within variance (σ_W^2) and the interaction variance (σ_{AB}^2). Dividing this mean square by the within variance, we get *F* for the interaction variance.

In a mixed model having *B* fixed and *A* random, the expected mean square for *A* contains the within variance (σ_W^2) and the variance due to *A* (σ_A^2). Hence, we divide this expected mean square by the within variance and obtain the *F* for *A*. The expected mean square for *B* contains three variances, namely, the within variance, the interaction variance and the variance due to *B*. The expected mean square for *A* × *B* interaction contains the above first two variances. Hence, the expected mean square for *B* is divided by the interaction variance. As the expected mean square for *A* × *B* interaction contains the within variance as well as the interaction variance, this mean square is divided by the within variance. Thus three *F*s in a mixed model are calculated as shown below.

Mixed model (A fixed, B random)

$$F \text{ for } A = \frac{MS \text{ for } A}{MS \text{ for interaction}}$$

$$F \text{ for } B = \frac{MS \text{ for } B}{MS \text{ for within-groups}}$$

$$F \text{ for } A \times B = \frac{MS \text{ for interaction}}{MS \text{ for within-groups}}$$

Mixed model (B fixed, A random)

$$F \text{ for } A = \frac{MS \text{ for } A}{MS \text{ for within-groups}}$$

$$F \text{ for } B = \frac{MS \text{ for } B}{MS \text{ for interaction}}$$

$$F \text{ for } A \times B = \frac{MS \text{ for interaction}}{MS \text{ for within-groups}}$$

PROBLEM OF CREATING EQUIVALENT GROUPS IN BETWEEN-SUBJECTS DESIGN

As we know in between-subject designs, a subject or participant is asked to participate in just one of the conditions of the experiment. Such design usually becomes necessary when subject variables (such as gender) are being studied or participating in one condition of the experiment tends to introduce changes in subjects in such a way that they can't be used in another condition.

When the experimenter uses the between-subjects designs, the major difficulty is found in creating groups that are equivalent to each other on all factors except for the IV. How to tackle this problem? There are two common techniques for creating equivalent groups in the between-subjects experiment. One technique is to use *random assignment* and another is to use *matching*.

As discussed earlier too, random assignment that is different from random selection of subjects, is a method for placing the selected subjects into different groups in such a way that every subject volunteering for the study has an equal chance of being placed in each group to be formed. The major goal of random assignment is clearly to take into account individual difference factors that may bias the study and spread them evenly throughout the different groups of participants. When large numbers of subjects are taken, it is, then, highly probable that random assignment will work to create equivalent groups. For completing a random assignment

of subjects to conditions in a way that guarantees an equal number of participants per group, the experimenter can use *block randomization*, which is defined as a procedure that ensures that each condition of the experiment has a subject randomly assigned to it before any condition is repeated a second time. In fact, here, each block contains all the conditions of the experiment in a randomized order.

Matching is another technique through which equivalent groups can be created. In matching, the participants are paired together on some trait and then, one member of each pair is randomly assigned to one of the group. Matching is often used in the following conditions:

(a) When the number of the subjects is small, so random assignment becomes risky and might yield unequal groups

(b) When the matching variable is anticipated to affect the results in some predictable fashion (that is, matching variable tends to correlate with dependent variable)

(c) When there is a reasonable way of assessing subjects/ participants on the matching variable

Of these three criteria, the first is true most of the time (because matching sometimes occur, even with larger *N*) but the second and third are essential.

WITHIN-SUBJECTS DESIGN

Till now we were discussing the between-subjects design because in the randomized-groups design, the matched-groups design and the factorial design two or more values of the independent variables are selected and one value is assigned to each group in the experiment. Thus in a between-subjects design, the experimenter compares the dependent variable score between two or more groups who are given differential treatment. In a within-subjects design the same group of subjects is treated differently in different experimental conditions and finally, their dependent variable scores are compared. This is also called *repeated-treatments design* because the same individuals are treated differently at different times and we compare their scores as a result of different experimental treatments. Let us illustrate this design with an example. Suppose using the within-subjects design, the experimenter wants to investigate the effect of a drug upon retention of a verbal task. He may test a group of subjects under two conditions, namely, in normal condition when the group has not taken a drug and a condition when it is under the influence of a drug. Thus each subject of the group has a pair of measures on the dependent variable. Subsequently, we may test the significance of the difference between means through appropriate statistical tests. Any change in the dependent variable measures may then be ascribed to the application of the drug. The above experiment can also be conducted through the between-subjects design where the experimenter is required to have two equivalent groups of subjects and he may assign one group to the experimental condition where the subjects will be under the influence of a drug and another group to the control condition where the subjects will be in a normal state, that is, without any influence of a drug. Subsequently, the mean scores obtained on the dependent variable measure may be compared and their significance of difference may be tested through appropriate statistical tests. The experiment based upon the within-subjects design is common in the field of learning, memory and psychophysics.

The within-subjects design may, for convenience, be divided into two parts: a design having two conditions and many subjects, and a design having more than two conditions and many subjects. In the former case, the same group is tested under two conditions only. The mean difference under these two conditions may be tested through the matched *t* test or Sandler's *A* test. The experiment described above illustrates the within-groups design having two conditions. The within-groups design may also be extended to a situation where more than two experimental conditions are utilized. The experiments based upon such a design have been frequently reported in the field of proactive inhibition. The experiment may be, here, designed as having five experimental conditions and the same group of subjects may be tested under each of the five conditions. In condition I the group may be required to learn zero list; in condition II it may be required to learn three lists; in condition III four lists; in condition IV five lists; and in condition V seven lists. At the end of each condition, the group may be required to learn one new list. After a time interval of one hour since learning this new list in each condition, the group may be asked to recall items from this new list. It is expected that the number of prior lists learned will proactively interfere with the retention of the new list. Greenberg & Underwood (1950) independently conducted several experiments using such within-groups designs on proactive inhibition and have shown that the greater the number of prior lists learned, the higher is the amount of proactive inhibition.

PROBLEM OF CONTROLLING SEQUENCE EFFECTS IN WITHIN-SUBJECTS DESIGN

As mentioned earlier, within-subjects design is one where each subject participates in all of the conditions of experiment. For such design, participating in one condition might affect how subjects tend to behave in other conditions. As a consequence, sequence or order effects can occur. Sequence effects include both *progressive effects* and *carryover effects.*

A within-subjects design becomes necessary when subjects are scarce. Besides, such design eliminates creating equivalent groups that occur with between-subjects designs.

Now the problem is how to control sequence effects in within-subjects design. Sequence effects include progressive effects and carryover effects. In an experiment using within-subjects design, trial No. 1 might affect the subject in someway so that performance on trial 2 may be steadily improved. Sometimes, repeated trials may also produce gradual fatigue or boredom and the performance may decline from trial 1 to trial 2. These two effects are called *progressive effects* since it is assumed that performance changes progressively from trial to trial. Likewise, in an experiment having two basic conditions such as A and B, experiencing condition A before B might affect the person differently than experiencing B before A.

An important way to control sequence effects is called as *counterbalancing* where more than one sequence is used. There are *two* general categories of counterbalancing depending upon whether participants or subjects are tested in each experimental condition just one time or are tested *more than* once per condition. A discussion follows.

1. When participants are tested once per condition

In some experiments, participants are tested in each of the conditions but tested only once per condition. Whenever the participants are tested once per condition in a within-subjects design, one easy solution to the problem of sequence effects is to use *complete counterbalancing,* where every possible sequence is used at least once. The total number of sequences needed is determined by calculating $X!$, where X is the number of conditions and '!' means mathematical calculation of factorial. For example, in an experiment if there are three conditions, there can be six possible sequences that can be used.

$$X! = 3! = 3 \times 2 \times 1 = 6$$

The six sequence in a study with conditions A, B and C would be:

ABC BAC
ACB CAB
BCA CBA

The major problem with complete counterbalancing is that as the number of conditions increases, the possible sequences needed also increase exponentially. As stated above, there are 6 sequences needed for three conditions but simply adding a fourth condition, it creates a need for 24 sequences $(4 \times 3 \times 2 \times 1)$.

Partial counterbalancing is another way to solve the problem of sequence effects when subjects are tested once per condition in within-subjects design. Whenever a subset of the total number of sequence is used, it is referred to as *partial counterbalancing.* In such a situation, it is possible to neutralize the sequence effects by the following rules:

(a) Every value of independent variable (or condition) should occur the same number of times at each ordinal position.

(b) Every value of the independent variable (or condition) precedes and follows every other value an equal number of times.

Suppose that the experimenter has used four conditions (or independent values) such as A, B, C and D which would require 24 different sequences $(4 \times 3 \times 2 \times 1)$. Out of these 24 different sequences, the following four orders may be chosen to satisfy the above two-point rules:

A B D C
B C D A
C D B A
D A C B

In this selection, we find that each condition occurs once at each ordinal position and each condition precedes and follows every other an equal number of time. Summing up data for all subjects will neutralize the differential impact of sequential effects.

2. When participants are tested more than once per condition

There are experiments where it becomes necessary for subjects to experience each condition more than one time. It often occurs in experiments in the field of sensation and perception. Suppose in an experiment on Muller-Lyer illusion, the experimenter is interested in measuring the extent of such illusion if the lines are presented in four different conditions: A, B, C, D.

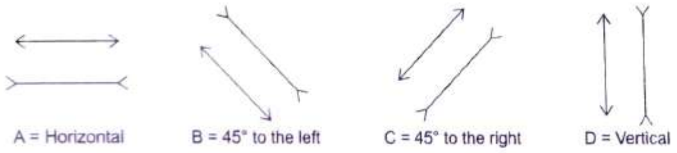

Figure 21.10 Set of four Muller-Lyer illusions: horizontal, 45° to the left, 45° to the right and vertical.

The sequence effects occurring in this experiment may be controlled by *reverse counterbalancing* and *block randomization.* In reverse conditioning the experimenter simply presents the conditions in one order, then presents them again in reverse order. In the experiment on illusions, the order of the reverse conditioning would be A–B–C–D–D–C–B–A. If the experimenter wants to have the subjects perform the task more than twice per condition, this sequence would be repeated as many times as necessary: One problem with reverse counterbalancing is that the participants can predict what is coming next.

Another way to present a sequence of conditions when each condition is presented more than once is that of *block randomization.* The procedure is also utilized in the context of how to assign subjects randomly to groups in between-group experiment. When block randomization is used in within-groups design, it is then ensured that every condition occurs once before any condition is repeated a second time. Within each block, the order of conditions is randomized and this strategy eliminates the possibility that the subjects can predict which condition will be coming as it happens in reverse conditioning. In the above experiment on illusion, the participants would encounter all four conditions in a randomized order, then all four again but in a block with a new randomized order, and so on for as many blocks of four as needed. A block randomization technique may produce either of the following two sequences:

C A B D A B D C
B C D A C A D B

These various counterbalancing techniques help in reducing sequence effects but some may do so imperfectly because counterbalancing technique assumes that the sequential effects are *linear* and this may not always be true, especially for carryover effects. Such carryover effects may create problems that counterbalancing might not cure. Suppose, for example, that the participants are required to learn two mazes—Maze A and Maze B. Learning of Maze A gives participants an insight about learning mazes in general but the same insight does not result when solving Maze B. If this really happens then in the sequence Maze A → Maze B learning A first will produce a lot of positive transfer that carries over to the learning maze B. On the other hand, learning maze B first will not produce any such positive transfer to learning of maze A in the sequence B → A. It means that the two sequences would produce what Poulton (1982) has called as *asymmetric transfer*. This obviously means that one sequence will produce a particular outcome that is not matched by the counterbalance sequence. If the experimenter expects asymmetric transfer, it would be wise to switch to a between-subjects design if such change is feasible.

COMPARISON OF BETWEEN-SUBJECTS DESIGN AND WITHIN-SUBJECTS DESIGN

As we have discussed in detail both the within-subjects design as well as the between-subjects design, we are in a position to make a comparative evaluation of these two designs.

1. In a within-subjects design the same group is used under all experimental conditions whereas in a between-subjects design, a separate group serves under each experimental condition. Obviously, then, there is the advantage of economy of subjects in a within-subjects design. Since each subject participates in all experimental conditions in the within-groups design, the data for all subjects completing the experiment are available in perfectly valid form.

2. Since the experimenter repeats the measures on the same group of subjects in a within-groups design, the factor of individual difference is automatically controlled. This increases the correlation among measures which, in turn, reduces the error variance. But in the between-groups design, the factor of individual difference is not controlled as different groups of subjects participate in different experimental conditions. This has the unfortunate effect of increasing the error variance.

3. If the experimental situation is such in which preparation for the experiment requires a fair amount of time and patience (mostly true in the case of neuropsychological and psychophysiological experiments), the within-groups design is preferred to the between-groups design because the same group of subjects with all its initial preparation, can more easily be tested under different conditions of the experiment than different groups of subjects in different conditions. For example, suppose the problem of the experiment requires that electrodes be fitted in the hypothalamus of animal subjects after a surgical operation, naturally, then, one would prefer that the design should be such that the same animals may work under all the experimental conditions so that the problems arising out of the necessary preparation can be minimized.

4. There are some experiments in which the within-subjects design should not be used. When the experiment is such that the order of presentation of conditions is likely to have a profound effect upon the dependent variable and there are obvious practice and fatigue effects, the within-groups design should not be preferred. In other words, when the experimenter expects a differential transfer effect from one experimental condition to another, he should not use the within-groups design. A differential transfer refers to the fact that the experimental treatment produces different effects when the experimenter moves from condition *A* to *B* than when he moves from *B* to *A*. This is also known as *asymmetric transfer* (Poulton 1982). Poulton is one of the vehement critics of within-groups design on the basis of asymmetric transfer. Let us illustrate this

with an example. Suppose the experimenter is conducting an experiment to examine the effect of knowledge of results (*KR*) upon a sensory-motor task such as a line-drawing task using a within-groups design. Further suppose that in order to counterbalance the effect of practice and fatigue, he uses the *ABBA* order. If he introduces first the *KR* condition in10 trials followed by the without knowledge of result (*WKR*) in the next 20 trials and again the *KR* condition in next 10 trials, there is possibility of unnecessary gain on the part of subjects, which may make the results unreliable. In the *KR* condition, the subjects are given the knowledge of the actual length drawn by them each time. With a practice of 10 trials they may develop some skills and special cues, which may help in a line-drawing task. Subsequently, when the *WKR* condition is started where they are not given the knowledge of the length drawn, they are likely to profit by their previous skills and special cues. This may have the effect of leading to a conclusion that there is no difference between the *KR* condition and the *WKR* condition. Obviously, then, this would be an unreliable conclusion undermining the importance of the knowledge of results. If instead, the between-groups design is used in which one group is given *KR* and another group is *WKR*, it is likely that the performance of these two groups would differ significantly. Probably, the group receiving the *KR* treatment would be more accurate in line-drawing task than the group receiving the *WKR* treatment.

Although between-subjects designs are considered more powerful, there are four reasons of choosing a within-subjects design.

(i) When only a smaller number of subjects are available, within-subjects designs are preferred.

(ii) Within-subjects designs are especially useful when no good matching tasks exist with which to equate the several independent groups of subjects.

(iii) Within-subjects designs possess comparatively a greater degree of convenience or efficiency (Posner 1973).

(iv) Within-subjects designs are especially needed for those areas of psychological research which require studying changes in subjects' behaviour over time. Panel or longitudinal studies and areas like psychophysics and scaling are most appropriate for such designs.

However, within-groups designs are not considered appropriate to study subject variables like age, sex and the like because such subject characteristics can't be successfully balanced.

EXPERIMENTAL DESIGN BASED UPON THE CAMPBELL AND STANLEY CLASSIFICATION

Apart from the experimental designs (excluding factorial design) discussed above, Campbell & Stanley (1963) have discussed 16 designs ranging from the poorest to very strong ones, which have proved very useful in psychological and educational researches. It is not possible to discuss all the 16 designs here. However, some of the important designs will be selected and discussed. For a detailed discussion readers are referred to Campbell & Stanley (1963).

In discussing their 16 experimental designs, Campbell & Stanley have used some symbols with which a reader is expected to be acquainted.

R: Random selection of subjects or random assignment of treatment to experimental groups

X: Treatment or experimental variable which is manipulated. When treatments are compared, they are labelled as X_1, X_2, X_3, and so on.

O: Observation or measurement or test. Where there is more than one *O*, an arbitrar subscript O_1, O_2, O_3, and so on, is used.

When one or more *X* and *O* occur in the same row, they indicate that these are being applie to the same persons. The left–to–right dimensions in which *X* and *O* occur, indicate tempora order and when *X* and *O* occur in vertical order to each other, they indicate that these two ar simultaneous. The parallel rows of symbols unseparated by a dashed line indicate that grou have been equated by randomization and when separated by dashed lines, they indicate that groups have not been equated by randomization.

PRE-EXPERIMENTAL DESIGN (NONDESIGNS)

There are three designs which actually do not qualify for the experimental designs because they do not provide a control group or the equivalent of a control group. Such designs do not control adequately threats of internal validity. These designs are called *pre-experimental designs* because they incorporate the least basic elements of an experimental design. Because these designs are inadequate in themselves, they are also sometimes referred to as nondesigns. The following are three pre-experimental designs.

One-Shot Case Study

The one-shot case study may be diagrammed as indicated below.

$$X\ O$$

As its name implies, in a one-shot case study the treatment X is given to a single group and subsequently, an observation O is made to assess the effects of treatment upon the group. This design has two important limitations. One is that it does not provide a control group and another is that it does not give any information regarding the members who are given the treatment. In view of these limitations, there is little justification for concluding that X caused O. Let us take an example. Suppose the principal of a college introduces the practice of giving monetary reward (X) to students who regularly attend their classes. After this practice has been in operation for a year, the principal observes that the students attend their classes regularly and disruptive activities in the classroom are minimized (O). On this basis the principal concludes that with the practice of giving monetary reward, the absenteeism and the disruptive activities in the classroom are reduced. This conclusion is, however, dubious because the principal does not know (a) whether or not factors other than monetary reward have contributed to the observed change in behaviour; and (b) whether there was a real change in the observed behaviour relative to their past behaviour.

One-Group Pretest-Posttest Design

This design is an improvement over the above design because the effects of treatment (X) are judged by making a comparison between pretest and posttest scores. However, no control group is used in this design. The design may be diagrammed as shown below.

$$O_1\ X\ O_2$$

For example, suppose the principal of a college wants to study the effect of movies in changing the attitude of a group of students. He first obtains some initial measures of attitude (O_1) and then, for an hour the students may be asked to see the film (X), which intends to bring a change in attitude. Subsequently, a measure of attitude change may be obtained (O_2). This design provides some information about two extraneous variables (that is, selection and experimental mortality), which are likely to endanger the internal validity of the experiment. The pretest scores indicate the initial state of the selected subjects and the posttest scores indicate the state of the experimental mortality of the subjects. However, the extraneous variables or the sources of internal invalidity like history, maturation, testing and instrumentation are not controlled by this design.

Static-Group Comparison (or Intact-group Comparison)

In this design, two groups are taken. One group (O_1) experiences the experimental treatment (X) and another group does not experience the experimental treatment (O_2). Subsequently, these two groups are compared. The design may be diagrammed as shown below.

$$\begin{array}{c} X\ O_1 \\ \cdots\cdots\cdots \\ O_2 \end{array}$$

Thus in this design, a control group (the group receiving no treatment) is used as a source of comparison for the treatment-receiving group or the experimental group. Because a control group is used, sources of internal invalidity like history, testing and instrumentation are controlled. The factor of statistical regression is also controlled. One of the main demerits of this design is that subjects of the control and the experimental groups are neither selected at random nor are they assigned randomly to the groups. Thus two groups are not equivalent (as there are no pretest scores) and hence the factor of selection is not controlled. As the samples are not randomly drawn, the interaction effects of selection (a source of external invalidity) are also not controlled. The dashed line indicates that the control group (O_2) and the experimental group (O_1) have not been equated by randomization. In fact, these three pre-experimental designs are examples of how not to do research if they are alternatives.

TRUE EXPERIMENTAL DESIGNS

There are three experimental designs, which are called true experimental designs. In these designs, the control group and the experimental groups are formed and their equivalence is established through randomization. These designs are called true experimental designs because all the factors or variables contributing to internal invalidity are controlled. Although true experimental designs are the strongest type of designs, in some situations it is difficult to conduct experiments based upon such designs. Three true experimental designs are presented as mentioned below.

Posttest Only, Equivalent-Group Design

This design is the most effective and useful true experimental design, which minimizes the threats to the experimental validity. The design can be diagrammed as shown below.

$$\begin{array}{ccc} R & X & O_1 \\ R & & O_2 \end{array}$$

In the above design, there are two groups. One group is given the treatment (X), usually called the experimental group, and the other group is not given any treatment called the control group.The use of a control group automatically controls the two extraneous variables, namely, history and maturation. Both groups are formed on the basis of random assignment of the subjects and hence, they are equivalent (there is no dashed line). Not only that, subjects of both the groups are initially randomly drawn from the population (R). This fact controls for selection and experimental mortality. Besides these, in this design no pretest is needed for either group, which saves time and money. As both the groups are tested after the experimental group has received the treatment, the most appropriate statistical tests would be those tests, which make a comparison between the mean of O_1 and O_2. Thus either *t* test or ANOVA is used as the appropriate statistical test.

Let us take an example. Suppose the experimenter, with the help of the table of random numbers, selects 50 students out of a total of 500 students. Subsequently, these 50 students are randomly assigned to two groups. The experimenter is interested in evaluating the effect of punishment over retention of a verbal task. The hypothesis is that punishment enhances the retention score. One group is given punishment (X) while learning a task, and another group receives no such punishment while learning a task. Subsequently, both groups are given the test of retention. A simple comparison of mean retention scores of the two groups, either through the *t* test or ANOVA, would provide the basis for refuting or accepting the hypothesis.

Pretest-Posttest Control Group Design

This design is similar to the previous one except for the fact that it also makes a provision for pretest for both groups before experimental and control treatments are administered. The design may be diagrammed as shown below.

$$\begin{array}{ccc} RO_1 & X & O_2 \\ RO_3 & & O_4 \end{array}$$

It is obvious from the above diagram that the design has two groups. One group receives the treatment (*X*) and another group receives no such treatment. By use of the control group this design controls some sources of internal invalidity like history, maturation and statistical regression. As subjects are randomly assigned to the control group and the experimental group, the variables like selection and experimental mortality, posing threats to internal validity, are also controlled. It is also obvious from the above diagram that both groups are given a pretest and a posttest. The very element of pretest, however, introduces one basic limitation. In this design there is no control over the gain on the posttest due to the experience on the pretest (called testing effect), which may reduce the internal validity of the experiment. Not only this, there is no control over the possible sensitization to the treatment, which a subject might develop due to pretest experience, thus reducing the external validity of the experiment. In other words, it can be said that interaction of testing and treatment (a source of external invalidity) is not controlled. The data in the pretest-posttest control group design can be statistically analyzed by making a comparison of gain scores for the control group and the experimental group. In other words, mean gain score of $O_2 - O_1$ can be compared with the mean gain score of $O_4 - O_3$ so that it can be easily ascertained whether or not the treatment has a differential effect on the groups. If the groups are wholly equivalent, the posttest means, that is, O_2 and O_4, can also be compared for ascertaining the impact of treatment upon the groups.

Solomon Four-Group Design

The Solomon four-group design developed by Solomon (1949) is really a combination of the two equivalent-groups design described above, namely, the posttest-only design and pretest-posttest design, and represents the first direct attempt to control the threats of the external validity. In this design, subjects are randomly assigned to two control groups and two experimental groups. Only one experimental group and one control group receive a pretest. All four groups receive a posttest. This design may be diagrammed as shown below.

$$\begin{array}{llll} R & O_1 & X & O_2 \\ R & O_3 & & O_4 \\ R & & X & O_5 \\ R & & & O_6 \end{array}$$

It is clear from this diagram that in this design, four groups are randomly set by the experimenter. As a matter of fact, in this design, two simultaneous experiments are conducted and hence, the advantages of replication are available here. The effect of *X* (treatment) is replicated in four ways: $O_2 > O_1; O_2 > O_4; O_5 > O_6$ and $O_4 > O_3$. This design makes it possible to evaluate the main effects of testing as well as the reactive effect (or interaction) of testing, maturation and history, thus increasing the external validity or generalizability. The factorial analysis of variance can be used as the appropriate statistical test. Because the design is complex from the methodological as well as the statistical point of view, it is less preferred to the above two true experimental designs.

Latin Square Design

This experimental design is used to examine whether the order or sequence in which subjects or participants receive multiple versions of the treatment has an effect upon the dependent variable. Thus when a researcher is interested in how several treatments given in different sequences or time orders affect a dependent variable, a Latin square design is used. For example, suppose a teacher of geography has three fundamental units to teach students: *how to use a compass, map reading* and *longitude-latitude* (LL) system. These units may be taught in any order but the teacher may be interested in knowing which order proves to be most beneficial to students for learning. In one class the order of teaching may be—map reading, then using compass and then the LL system. In another class the order may be using compass, map reading and then LL system. Still in another class the order may be LL system, then compass usage and then map reading. The teacher takes examination after each unit and students take a comprehensive examination at the end of the entire term. Since the students are randomly assigned to three classes, the teacher can easily determine whether or not presenting units in specific order resulted in improved learning.

QUASI-EXPERIMENTAL DESIGNS

A quasi-experiment is one that applies an experimental interpretation to results that do not meet all the requirements of a true experiment. It means when the situation is such that the experimenter has some control over the manipulation of independent variables but fails to arrange for the other basic requirement of a true experiment, that is, creating equivalent groups. Thus, a quasi-experiment is basically an attempt to *simulate* the true experiment and, therefore, has also been called a *compromise design*. Quasi-experimental designs are partly like true experimental designs. They control some but not all extraneous variables, which give threats to the internal validity of the experiment. The quasi-experimental designs provide control over when and whom the measurement is applied but as subjects are not randomly assigned to the experimental and the control group, the equivalence of the groups is not maintained, and thus, it leaves some uncontrolled threats for validity of the experiment. In such designs we don't truly manipulate the independent variable. Such variables are called 'quasi-independent variables' and studies that employ them are called quasi-experiments where we lay out the design and compare the scores between the different conditions as in a true experiment, but we only appear to administer the variable. In a quasi-experiment, the researcher cannot randomly assign subjects to be exposed to a particular condition. Instead, subjects are assigned to a particular condition because they already qualify for that condition due to some inherent characteristics. Age, sex, race, background experiences or personality characteristics are some of the examples of quasi-independent variables. Such designs are better than pre-experimental designs, although they are not as adequate as the true experimental designs. In a way, then, the quasi-experimental designs are somewhere in between the pre-experimental designs and the true-experimental designs. Campbell & Stanley (1963) have discussed several types (excluding the variants of some types) of quasi-experimental designs, some of which are presented below.

Time-Series Design

Sometimes it happens that a control group or a comparison group cannot be included in an experiment because of the situation in which the experiment is being conducted. Still, the experimenter wants to have a design, which may exercise a better control over the extraneous variables. The time-series design is one such design, which can be followed in the situation described above. This design can be diagrammed as indicated below.

$$O_1 O_2 O_3 O_4 \quad X \quad O_5 O_6 O_7 O_8$$

It is obvious from the above diagram that a series of pretests are given to the group. Subsequently, the treatment (*X*) is given and a series of posttests are given to the same group. This design differs from the single group pretest-posttest design because instead of giving a single pretest and posttest, a series of pretests and posttests are given. The extraneous variables like maturation, testing, selection and experimental mortality are well controlled. However, a variable like history is not controlled because the subjects are exposed daily to the different kinds of stimulations beyond those under the control of the experimenter. In analyzing the results of the time-series design, simple statistical comparisons between O_4 and O_5 are avoided. However, comparisons among all other pairs of adjacent points are recommended.

Equivalent Time-Samples Design

The equivalent time-samples design is an extension of the time-series design with the repeated introduction of the treatment or the experimental variable. Like the time-series design, in equivalent time-samples design a single group is used and the group is exposed to repeated treatments in some systematic way. The design may be diagrammed as indicated below.

$$X_1O_1\quad X_0O_2\quad X_1O_3\quad X_0O_4$$

It is clear from the above diagram that in the equivalent time-samples design the treatment (X_1), instead of being given for a single time, is introduced and reintroduced with some other experience (X_0), which is available in the absence of the treatment. The variable history, which is the major limitation of the time-series design, is well controlled by presenting X on several separate occasions. Other extraneous variables (posing threats to internal validity) such as maturation, testing, instrumentation, regression, selection, and experimental mortality are also well controlled. Thus this design provides not only an extension but also an improvement over the time-series design. The design can be illustrated with the following example.

Suppose the experimenter wants to study the effect of showing a 25-minute pro-nationalization film on the attitude towards nationalization for a group of grade IX students. The experimenter shows the film to a group of students (X_1). After that, he administers the measure of attitude change towards nationalization (O_1). After a few days, the experimenter briefly discusses the general usefulness of nationalization with them in their class (X_0). Again, the test of their attitude can be made (O_2). After a few days, the group witnesses the same film (X_1) and the measure of attitude may be obtained (O_3). Following this, the experimenter discusses in detail every aspect of nationalization (X_0). Their attitude towards nationalization is again measured (O_4). The statistical comparison of O_1 and O_3 with O_2 and O_4 helps the experimenter to compare two experiences. The time effects are easily determined by comparing O_1 and O_2 to O_3 and O_4. Although the equivalent time-samples design is an improvement over the time-series design, it introduces two important problems of external validity. *First*, if the effect of treatment (X_1) after administering over the time is different from the effect when it is introduced and reintroduced, it will be difficult for the experimenter to make a generalization about the continuous effect of X_1 beyond the experiment. This has the effect of lessening the external validity of the experiment. *Second*, the repeated administrations of the treatments may lead to an adaptation which, in turn, may lessen the generalizability.

Nonequivalent Control Group Design

In psychological and educational research, often the experimenter is faced with a situation in which he has to work with intact groups (that is, groups whose membership is prefixed and cannot be altered by the experimenter). For example, the teacher may provide the experimenter two classes of geography but he may not agree with the latter in reconstituting them for experimental purposes. In such a situation, the experimenter has to accept them as intact groups. When reconstitution of subjects is not allowed, the experimenter cannot randomly assign them to the control group and the experimental group and hence, their equivalence is suspected. In working with such intact nonequivalent groups, the nonequivalent control group design is recommended. This design may be diagrammed as shown below.

$O_1\ X\ O_2$

……

O_3 O4

The above design is similar to the pretest-posttest control group design except that it does not bear the subscript *R*. This means that except for random assignment of subjects to the experimental and the control conditions, which occur in case of the pretest-posttest control group design, this design is identical to that design. As the experimenter is not allowed to randomly assign the groups to the experimental and the control condition, their equivalence is not granted. This creates difficulty in controlling a variable like selection. However, this difficulty can be overcome by comparing the intact groups on pretests, that is, on O_1 and O_3. For the satisfaction of the experimenter on the criterion of equivalence, the intact groups can be compared on any control variables relevant to selection and potentially relevant to the treatment such as age, sex, intelligence, socio-economic status, etc. The statistical analysis consists of comparing the mean gain score of the treatment group ($O_2 - O_1$) to the mean gain score of the nontreatment group ($O_4 - O_3$). The nonequivalent control group design, however, should not be used, where one of the intact groups consists of volunteers and the other nonvolunteers. This is because the volunteers are different from nonvolunteers and a comparison of the two does not provide for control of a variable like selection.

Consider an example to illustrate this design. Suppose the experimenter wants to know the effect of a week's training intended to improve the map-drawing skill among students of a geography class. For this purpose, the principal of a school provides the experimenter with two classes of geography, each consisting of 20 students. The principal does not permit reconstitution of these two groups in any way and requests the experimenter to handle them as intact groups. Following the request of the principal, the experimenter decides to use the nonequivalent control group. Which of these two groups would act as a control group and as an experimental group was decided randomly by flipping a coin by the experimenter. The experimenter, however, ascertained the equivalence of these two intact groups on age, sex, presence of physical disorders and so on, and was satisfied that these groups could be regarded as equivalent in all these respects. Both the groups were administered the map-drawing test as pretest measures. Subsequently, the experimental group was given training in map-drawing work (X) and the control group was not given any such training. After that, both groups were readministered the map-drawing (O_2 and O_4). The mean gain scores (posttest minus pretest) of the two groups were compared and it was found that the difference between $O_2 - O_1$ to the difference between $O_4 - O_3$ would be positive and statistically significant.

Counterbalanced Design

In counterbalanced designs, the experimental control is achieved by randomly applyin experimental treatments. Such designs are called *Crossover designs* (Cochran & Cox 1957; Co 1958), *Switch-over designs* (Kempthorne, 1952), and *Rotation experiments* by McCall (1923). The name 'counterbalanced designs' was, however, given by Underwood (1949). F counterbalancing the treatments, generally the Latin-square management in which eac treatment appears once and only once in each column and in each row, is made. counterbalancing design in which four treatments have been randomly given to four groups o four different occasions is diagrammed below.

Gr. A $X_1O\ X_2O\ X_3O\ X_4O$

……

Gr. B $X_2O\ X_4O\ X_1O\ X_3O$

……

Gr. C $X_3O\ X_1O\ X_4O\ X_2O$

……

Gr. D $X_4O\ X_3O\ X_2O\ X_1O$

……

Variables like history, maturation, testing, instrumentation, selection, regression, a experimental mortality (posing threats to internal validity) are well controlled by counterbalanced designs. One source of external invalidity, however, creeps into counterbalanced design, that is, multiple X interference lowers the generalizability.

Separate-Sample Pretest-Posttest Design

The separate-sample pretest-posttest design is specially suited to those situations in which the experimenter cannot assign treatments to all subjects at a time. Hence, he is forced to select a sample and administer the treatment. Then, again another sample is taken and the same treatment is repeated. Consider a situation in which there are 2,000 persons who are to be trained but the experimenter cannot train more than 200 persons at a time. In such a situation the training programme has to run continuously each time with a new set of persons and at the same time, the experimenter cannot withhold treatment from any person. Consequently, he cannot assign persons to training conditions and nontraining conditions. As such, no true experimental design can be applied here. To deal with such a situation, a separate-sample pretest-posttest design may be adopted. For this purpose, the experimenter decides to take a one-group pretest-posttest design and repeat it as shown below.

$$O_1 \quad X \quad O_2$$

……………………………………

$$O_3 \quad X \quad O_4$$

If the one-group pretest-posttest design is used without repetition, one may say that a variable like history is not controlled because the subjects may be influenced by several other events occurring simultaneously with the treatment (*X*) that might have produced O_2. But when that design is repeated thus constituting a separate-sample pretest-posttest design, the variable like history is controlled because it is less likely that the same events would have occurred simultaneously to *X* on both administrations. If O_2 exceeds O_1 and O_4 exceeds O_3, we have direct evidence for concluding this. Despite this, the separate-sample pretest-posttest design fails to control the three factors of internal invalidity such as testing effects, maturation and interaction of selection and maturation.

Patched-up Design

In a patched-up design the experimenter starts with an inadequate design and then, adds some features so that recurrent factors producing invalidity may be maximally controlled. The patched-up design shown below is a combination of two different pre-experimental designs, neither of which is adequate in itself, but which become adequate when combined. Usually, the experimenter selects such inadequate designs so that on combination their weaknesses are overcome and their strengths are increased. For example, let us take two pre-experimental designs, namely, one-group pretest-posttest design and the static-group comparison design. These two designs are inadequate in themselves because they fail to control many extraneous variables which threaten the internal validity. For example, the former design fails to control variables like history, maturation and statistical regression but controls selection, whereas the latter design fails to control selection but controls maturation and history. Hence, a patched-up design may be built on the basis of these two designs so that their mutual limitations may be overcome. The resulting patched-up design may be diagrammed as shown below.

Class *A*		X	O_1	
Class *B*			O_2	X $\quad O_3$

In the above diagram, the O_3 and O_2 comparison represents the one-group pretest-posttest comparison and controls variables like selection to a greater extent but fails to control history and maturation. Likewise, an O_2 and O_1 comparison is an intact-group comparison and controls maturation and history but fails to control selection. If O_3 versus O_2 and O_2 versus O comparisons can be regarded as equal then history, maturation and selection are automatically controlled so that the outcome cannot be accounted for in terms of these extraneous variables.

A simple illustration of a patched-up design given above reveals some interesting points about this design. In a patched-up design every subject gets the treatment and therefore, the experimenter cannot withhold treatment by assigning some subjects to the control group. The experimenter can, however, control when and to whom the treatment is given at a particular time during the course of experimentation. The essence of the patched-up design is that groups, like a separate sample pretest-posttest design, are tested in sequence and subsequently, they are compared.

Apart from the above quasi-design, there are also other quasi-independent variables that have been widely studied. One very widely studied quasi-independent variable in psychology is the passage of time. The field of developmental psychology is built around age, focusing on how it relates to various kinds of development in social, emotional and cognitive behaviour. Passage of time is also widely studied in different other settings where we compare experienced with inexperienced workers or study changes in memory abilities with advancement in age. These are quasi-independent variables because the experimenter cannot randomly assign people to be of a certain age or to experience only a certain amount of work. There are three general designs for studying the passage of time: *longitudinal design, cross-sectional* design and *cohort design*. These may be discussed as under.

Longitudinal Design

In longitudinal design, the researcher usually measures a group of subjects in order to observe the effect of passage of time. In fact, such designs are confounded by extraneous events that occur during the study and they may not generalize over time. The researcher may like to study vocabulary development in a group of female children by testing them yearly from ages five to ten. A diagram showing the plan for such study has been presented in Table 21.19.

Table 21.19 Diagram of longitudinal study showing repeated observations of each subject at different ages

	Age in years					
	5	6	7	8	9	10
Subject 1	*X*	*X*	*X*	*X*	*X*	*X*
Subject 2	*X*	*X*	*X*	*X*	*X*	*X*
Subject 3	*X*	*X*	*X*	*X*	*X*	*X*
: :	:	:	:	:	:	:
	$\overline{X}$	$\overline{X}$	$\overline{X}$	$\overline{X}$	$\overline{X}$	$\overline{X}$

In Table 21.19, *X* and $\overline{X}$ show the score and mean respectively. Adding vertically, any differences between the mean scores for the conditions reflect changes in vocabulary as a function of age.

Longitudinal study can also be done over a shorter period of time. For example, Nelson and Sutton (1990) conducted one study in which they examined white-collar workers over a nine-month period to determine how they coped with work-related stress.

Longitudinal study has some advantages and disadvantages. Important advantages are as follows.

(i) This keeps subject variables reasonably constant between the conditions. For example, in the above study, planned and shown in Table 21.19, the researcher is able to keep variables like genetic make-up, their parents, and so on, constant because the same children as they age, are being observed.

(ii) Such a study avoids problems associated with sample nonequivalence.

(iii) It permits the formulation of cause-and-effect statements with more certainty than other research designs.

(vi) Such a research design can portray growth increments and patterns.

Longitudinal design has the following disadvantages:

(i) Longitudinal study is inherently confounded by any extraneous variable that the subjects might experience during the entire period of study.

(ii) A very lengthy longitudinal study may not generalize adequately to future generations because the society and culture are regularly and constantly changing with time. Thus such a research design lacks temporal validity.

(iii) Since longitudinal study requires repeated measures, successive responses may easily be confounded by carryover effects or by response sets and the results may be misleading because of death and for dropouts during the study.

(iv) In such a research design, the researcher cannot control the environment of the subjects between testing periods.

(v) This study locks researchers into an earlier research design and theory.

(vi) This research design is expensive in time and money. It also jeopardizes all the previous expenditures of time and money if research funds give out.

Cross-Sectional Design

Cross-sectional design is a between-subjects quasi-experiment in which the researcher observes the subjects at different ages or at different points in temporal sequence. Thus the researcher may select a cross section of ages, testing the vocabulary of a group of 5-year olds, another group of 6-year olds, and so on. Templin (1957) provided us with a good illustration of a cross-sectional study. In fact, he investigated language development by selecting sixty children at a number of age levels—3, $3\frac{1}{2}$, 4, $4\frac{1}{2}$, 5, 6, 7 and 8 years. Table 21.20 presents a diagram of cross-sectional study in which observation of different groups of subjects from ages 5 to 10 with respect to vocabulary development has been shown.

Table 21.20 Diagram of cross-sectional study showing observation of subjects of different age groups

5	6	7	8	9	10
X	X	X	X	X	X
X	X	X	X	X	X
X	X	X	X	X	X
:	:	:	:	:	:
$\bar{X}$	$\bar{X}$	$\bar{X}$	$\bar{X}$	$\bar{X}$	$\bar{X}$

In Table 21.20, X represents vocabulary score and $\bar{X}$ represents mean of different age groups.

Cross-sectional design has also some advantages and disadvantages. The following are the major advantages.

(i) Cross-sectional study can be conducted rather quickly and easily. Thus it saves a great deal of time.

(ii) It costs less than longitudinal study.

(iii) Such a study demands no continuity on a long-term basis for obtaining cooperation among research workers.

(iv) Such a study does not require that the data be frozen over long periods until subjects reach the desired time for retesting.

The major disadvantages of cross-sectional study are as under:

(i) Such a study does not indicate the direction of change taking place within the sample groups.

(ii) In a cross-sectional study, subjects of the same chronological age but different maturational age are lumped together. Such averaging may conceal many changes, especially those associated with the growth spurt.

(iii) Such a study neglects the continuity of development as it occurs in a single individual.

(iv) In cross-sectional study, comparability of the groups being studied is always uncertain. The researcher can never be sure that the reported age related differences between subjects are really not the product of other differences between the groups, namely, differences in intelligence, diet or sociocultural environment. This confounding is called *cohort effect* which occurs when age differences are confounded by differences in subject history. Thus, the comparability of the groups can only be assumed. Several development psychologists have demonstrated that this assumption is frequently invalid (Baltes 1968; Woodruff and Birren 1972).

Despite these disadvantages, cross-sectional designs provide an immediate snapshot comparison of subjects who differ in age or time-related experiences. To the extent the various groups match across conditions and eliminate confoundings in a cross-sectional design, the researcher can determine easily the impact of temporal differences upon dependent variables.

Cohort Design

In a cohort design the researcher conducts a longitudinal study of several groups, each from a different generation. Suppose for example, the researcher studies vocabulary development in one generation of children when they are five years old in 1990, in another generation of children beginning when they are of five years in 1995, and in another generation of children beginning when they are of five years in 2000. This design has been set up and analyzed in Table 21.21.

Table 21.21 Diagram of a cohort study repeated measures over ages

		5	6	7	8	9	10		
1990 Subjects	1	X	X	X	X	X	X	$\bar{X}$	
	2	X	X	X	X	X	X		
	3	:	:	:	:	:	:		
1995 Subjects	1	X	X	X	X	X	X	$\bar{X}$	Generation main effect
	2	X	X	X	X	X	X		
	3	:	:	:	:	:	:		
2000 Subjects	1	X	X	X	X	X	X	$\bar{X}$	
	2	X	X	X	X	X	X		
	3	:	:	:	:	:	:		
Age main effect		$\bar{X}$	$\bar{X}$	$\bar{X}$	$\bar{X}$	$\bar{X}$	$\bar{X}$		

In Table 21.21, X represents vocabulary score and $\bar{X}$ represents mean of scores. The columns denote the repeated testing of a child at different ages beginning from 5 to 10 years, and the three rows represent three respective generations. Collapsing vertically over the three generations, the mean scores for each age are obtained and the difference between the means demonstrate the main effect of different ages irrespective of the subject's generation. In fact, this part of the operation provides knowledge about longitudinal or developmental information. On the other hand, when we add scores horizontally, the researcher gets information about the main effect mean of each generation, regardless of age. If no significant mean differences are found, the researcher does not get any evidence that scores differ on the basis of the subject's generation. However, if there occurs differences between these three groups then *cohort effect* is said to be present and the developmental results are less likely to generalize to other generations. Likewise, there may be interaction between age and generation. If there is absence of significant interaction between age and generation, it would mean that changes in scores as a function of age are basically similar or *parallel* irrespective of each generation's own uniqueness and history. However, a significant interaction would indicate that type of changes visualized with age depends on which generation the researcher is studying.

Cohort design possesses advantages and disadvantages more or less similar to longitudinal design.

EX POST FACTO DESIGN

The reader has already been introduced to the ex post facto experiment discussed in Chapter 15. In ex post facto research, it can be said that the experimenter, instead of creating a treatment, evaluates the effects of a naturalistically occurring treatment after that treatment has occurred. He tries to relate the outcome (or the dependent variable measure) with already occurred treatments. In ex post facto research, the researcher gives the treatment not by manipulation but by selection. Because the independent variables are handled by selection in *ex post facto* research, sometimes it is difficult to find out the cause–effect relationship between the dependent variable and the independent variable.

There are two common types of *ex post facto* design, namely, *correlational design* and *criterion-group design*.

Correlational Design

A correlational approach (also known as the psychometric approach) is one in which the experimenter collects two or more sets of data from the same group of subjects so that the relationship between the two subsequent sets of data can be determined. While the correlations are the common statistics which are employed in analysis of these data, other statistics commonly regarded as variants of correlational techniques, can also be utilized. The correlational design may be diagrammed as follows:

$$O_1 \qquad O_2$$

Suppose the researcher wants to investigate the relationship between the intelligence and problem-solving ability of a group of children randomly taken from Class VI. For this purpose, the experimenter will administer the measure (or test) of intelligence (O_1) and subsequently, a test of problem-solving ability will be administered (O_2). Thus the researcher will have two sets of data before him. He may apply appropriate correlational techniques depending upon the nature of the data. If the obtained correlation coefficient is positive and significant, the researcher may conclude that the higher the intelligence, the greater the ability to solve a problem. In fact, a strong and significant relationship between O_1 and O_2 suggests one of three possible meanings.

1. The variable measured by O_1 has caused O_2.
2. The variable measured by O_2 has caused O_1.
3. A third or unmeasured variable has caused both O_1 and O_2.

But as the experimenter has not manipulated the variable, it is difficult to say which of the above three interpretations accounts for the obtained relationship. Whenever there occurs a weak relationship between O_1 and O_2, all the above three meanings are rejected. The correlational design, thus, cannot be regarded as an adequate design for establishing a cause–effect relationship among variables.

Criterion-Group Design

In the criterion-group design, as its name implies, the experimenter tries to ascertain what has caused the particular state of condition by contrasting the characteristics of the group which possesses the criterion behaviour with those who do not. The criterion-group design may be diagrammed as shown below.

$$\begin{array}{ccccccc} C & O_1 & O_1 & C & O_2 & C & O_1 \\ & \text{or} & \ldots\ldots\ldots & & \text{or} & \ldots\ldots & \\ & O_2 & O_3 & & O_4 & & O_2 \end{array}$$

In the above diagram, the letter C indicates the selection of an experience according to the criterion fixed by the experimenter. A close scrutiny of the design reveals that this is very similar to the correlational design and pre-experimental design, particularly the static-group comparison design. The criterion-group design may, therefore, be used in situations where intact groups of criterion and noncriterion subjects are available or where subjects with criterion experiences are randomly selected from a larger group of subjects all qualifying on the criterion variable. Let us take an example. Suppose the experimenter is interested in knowing the origins of divergent thinking among a group of 50 students randomly selected from Class IX. In other words, the purpose of the experimenter is to explore what are the factors, experience, types of personality, etc., that may cause divergent thinking among the children. As a first step, the experimenter will administer a test of divergent thinking to the group of students. On the basis of the test scores, he would identify two criterion groups—one group consisting of those students who score higher on the test and another group consisting of those students who score lower on the test. Subsequently, the experimenter may try to find out the general environment into which they were reared, the attitudes and education of their parents, etc., by taking a structured interview of their parents. Because the information or experiences gained by the experimenter on the basis of the interview have not been manipulated in any way, it is difficult for the experimenter to conclude that a particular set of experiences has caused the divergent thinking. However, this criterion-group study definitely provides some testable hypotheses in terms of the potential causes, which by using a quasi-experimental design or true experimental design may be tested in a more scientific way.

STEPS IN EXPERIMENTATION

When the problem has been selected, the experimenter plans to conduct an experiment. The following are the main steps in experimentation.

1. *Label the experiment*: As a first step, the experimenter should clearly label the title of the experiment. Not only this, he should also determine the time and location of the experiment.

2. *Review of the literature*: The review of the literature is very relevant to the experiment and is a very important aspect of the experimental plan. There are four reasons why a review by the experimenter is essential. *First*, it helps the experimenter in formulating an appropriate problem. *Second*, it also helps the experimenter in discarding any vague notion regarding a problem, already existing in his mind so that he can modify that problem to the extent of conducting a good experiment. *Third*, the review of the literature helps the experimenter in avoiding unnecessary duplication of a problem already well studied. If the problem finally selected by the experimenter is such that it has already been well investigated by somebody else, there seems no sense in replicating the same experiment unless the experimenter aims at confirming the findings of the earlier experiment. *Fourth*, the review helps the experimenter in acknowledging the ways in which the extraneous variables can be controlled. Thus, the review of the literature is an important step in the experimental plan. The most common source of review is journals and books. For *psychological researches* and educational researches, the psychological abstract is also one important such source.

3. *Formulation of the problem*: Generally, the experimenter decides to conduct an experiment in a field because there is a lack of scientific knowledge in that field. By formulating a problem, he expresses this lack of knowledge. The experimenter formulates a problem in an unambiguous manner, in short sentences, preferably in question form.

4. *Formulation of the hypothesis and defining the variables:* The next step in an experimental plan is to formulate the hypothesis. Usually, the dependent variable and the independent variable are first identified and then defined, and subsequently, the tentative relationship between the two is stated in the form of a sentence (called the *hypothesis*). The "if .....then...." relationship is the most common form of expressing a hypothesis. Apart from expressing a hypothesis in a common language, it can also be expressed in terms of mathematical and logical expression.

5. *Preliminaries and apparatus*: Preliminaries include a brief history of the subjects, including their name, age-range, sex, educational status, etc., although some researchers have objected to these items of information. Apparatus is the commonly used tools in psychological experimentation. In an experimental set-up, two major functions are served. First, apparatus helps the experimenter in proper administration of the experimental treatment, and second, it helps the experimenter in making a careful record of the behavioural changes occurring as a function of manipulation of the independent variables. There are several types of apparatus, which are used in behavioural experimentations. Hence, it is very difficult to enlist all apparatus. However, in the field of learning and transfer of training, the memory drum, mirror-drawing board, card-sorting tray, and stylus maze board are the common items. Likewise, for studying geometrical illusion the Miller–Lyer illusion board is frequently utilized. For studying psychogalvanic responses, the psychogalvanometer is used. Sometimes, the apparatuses are not accurate and may not be properly calibrated. Hence, the experimenter should be cautious in selecting and using apparatus of any kind in the experiment.

6. *Control of extraneous variables*: The experimenter must clearly identify those extraneous variables, which are likely to influence the dependent variable. He must prepare a plan to control these variables so that the changes occurring in the dependent variables may be due only to the manipulation of the independent variables. At the beginning of this chapter, a special discussion has already been made regarding the control of extraneous variables. Sometimes, the experimenter is faced with a situation in which the nature of extraneous variables is such that they are beyond his control. In this type of situation, he makes an assumption that the relevant variables are equally influencing all the conditions of the experiment. If this assumption is tenable, he can proceed with it, but if it becomes tenuous, it is best dropped.

7. *Design of the experiment*: There are several designs which are used in behavioural researches. Some of them have already been discussed in the preceding section. The selection of the design is made keeping in view the general purpose and hypotheses of the experiment. If the purpose of the experiment is to make a simple comparative study between two groups, the experimenter may use the *two-randomized-groups* design. But if the purpose also involves the evalution of the effect of the interaction, he may prefer a *factorial design*.

8. *Selection of subjects*: At this step, the experimenter takes a decision regarding the modes of selection of subjects and the ways they may form the groups. There are three steps, which are directly concerned with selection and assignment of subjects to the different conditions of the experiment. At the first step, the subjects are randomly selected from a wider well-defined group. The wider group is commonly known as the population and may consist of inanimate objects or living organisms. The selected number of subjects is usually called a sample. The experimenter usually wishes to generalize the obtained results in regard to the wider group from which the samples are drawn. In order to achieve maximum generalization, he tries to select a representative sample and random selection of subjects is one way to maintain the representativeness of a sample. The details of the techniques of random selection have already been discussed in Chapter 14. When the subjects are selected from the population, the next step is to randomly assign them to different groups. Usually, the number of groups in any experiment is determined by the number of conditions. Let us suppose that the experimenter wants to conduct an experiment in which there are only two conditions. Hence, for this experiment there will be two groups. Let us further assume that there are 30 subjects available for the purpose. At this second step, these subjects may be randomly assigned to two groups. This may easily be done with the help of the table of random numbers or with simple flipping of coins. Thus, 30 subjects are randomly divided into two groups. Now, one of the groups is to be designated as the *experimental group* and the other as the *control group*. This constitutes the third step, which should also be done in a random way by flipping a coin. The experimenter may predetermine by flipping the coin that 'head' will be called the experimental group and 'tail' will be called the control group. In this way, we see that the entire process of selection and assignment of subjects to different conditions is carried out in a random way so that the experimenter can make a nearly perfect and sound generalization with the minimum possible errors.

9. *Procedure of the experiment*: At this step, the experimenter makes a detailed plan regarding the methods of giving treatments to subjects and recording the resulting data in data-collection tables. The experimenter also specifies the instructions to the subjects. When the experimental procedure is complex and lengthy, the experimenter may expect a few errors to creep into the procedure. He should first try out a pilot experiment, the experimental procedure of which might give some indication about those likely errors.

10. *Statistical treatment*: The obtained data in the experiment are evaluated by statistical tests. There are several types of statistical tests, some of which are appropriate to one type of experimental design and data and some are appropriate to another set of data and experimental design. Not only this, these statistical tests rest upon certain assumptions. Before applying any statistical test, it is reasonable to examine whether or not the obtained data satisfy those assumptions. A statistic applied ignoring its assumptions is likely to yield unsatisfactory results and hence, is not dependable. A detailed discussion of the common statistical tests appears in Chapter 23. Needless to say, lack of appropriate statistical tests or a test selected without much thinking and insight is likely to invalidate the results of the experiment.

11. *Preparation of the discussion report*: After statistically analyzing the data, a discussion report is prepared where the experimenter explicitly writes whether or not the hypothesis is supported by the obtained results. The experimenter also relates his findings with findings of researchers who have already done experiments relating to the problem. Finally, he concludes and summarizes his main findings under a separate heading.

12. *Generalization of the obtained findings*: The term 'generalization' here means that what is true for the sample is also true for the whole population. The experimenter wants to generalize his obtained findings for the whole population from which samples were randomly selected to represent it. As a matter of fact, generalization is, to a greater extent, dependent upon two factors. First, the factor of *specificity* by which is meant the extent to which the population has been well specified. A nonspecified population is vague and then, generalization may have little meaning for the population. The second is the factor of *representativeness of the sample*. If the sample has been selected in a manner which ensures maximum representativeness of the population, generalization is likely to be correct and accurate.

When the experiment has been conducted, data are analyzed and a conclusion is reached, the experimenter then writes up the report of the experiment. Writing a research or experiment report requires that the experimenter must have knowledge of a certain format according to which reports are prepared. Chapter 24 is entirely devoted to this aspect of research work. In general, it can be said that the conventions as laid down in the *Publication Manual* of the American Psychological Association are quite satisfactory ones and can safely be adopted for preparing a report.

☑ Review Questions

1. What do you mean by research design? Discuss the basic principles of experimental design.
2. What is meant by factorial design? Discuss the advantages and disadvantages of a factorial design.
3. Citing relevant examples, discuss the criteria of a research design.
4. Make a comparative study of the between-subjects design and the within-subjects design. Which of the two do you prefer for studying a social problem? Give reasons.
5. Distinguish between experimental design and quasi-experimental design with examples.
6. What reasons might a researcher choose in using a within-subjects design?
7. Distinguish between the complete within-subjects design and the incomplete within-subjects design.
8. How do subject variables differ from manipulated independent variables in experimental research?
9. Discuss the importance of manipulation, holding conditions constant and balancing in ensuring the internal validity of an experiment.
10. Distinguish between random selection and random assignment in terms of their uses in experimental researches.
11. Citing an example, make a comparative study of longitudinal design and cross-sectional design.
12. Discuss the basic features of cohort design.

◆

22

QUALITATIVE RESEARCH

CHAPTER PREVIEW

- Meaning and Essential Features of Qualitative Research
- A Qualitative Research Model: Five Components
- Relevance of Qualitative Research
- Brief History of Qualitative Research
- Themes of Qualitative Research
- Theoretical Perspectives of Qualitative Research
- Research Design Strategies of Qualitative Research
- Sampling Techniques of Qualitative Research
- Data Collection Techniques in Qualitative Research
- Data Analysis and Interpretation
- Comparison of Methods of Qualitative and Quantitative Data Analysis
- Combining Qualitative and Quantitative Approaches

Qualitative research is an ongoing process of proliferation with new methods and approaches appearing and it is being taken up by more and more disciplines apart from psychology, education and sociology as a core part of their curriculum. In fact, qualitative research has become particularly relevant in the last decades of the twentieth century and at the beginning of the twenty-first century. Let us discuss qualitative research in its seminal background. As we know, the research methodology can be divided into two major paradigms: *Logical-positivism* or *logical-positive paradigm* and *phenomenological paradigm*. Logical-positive paradigm is based upon the assumption of natural sciences and utilizes experimental research methodologies such as description of scientific method, hypothesis testing, etc. Quantitative research is based upon logical-positive paradigm. Qualitative research is based upon the phenomenological paradigm, which uses a variety of interpretive research methodologies.

MEANING AND ESSENTIAL FEATURES OF QUALITATIVE RESEARCH

Confusion exists over just what qualitative research is. There are several reasons behind it. *First,* qualitative research represents a negative connotation, that is, qualitative research is only what quantitative research isn't rather than positively saying what it is. *Second,* some researchers mistakenly think that qualitative research is a unitary approach whereas in reality, qualitative research consists of a variety of alternative approaches to the traditional, positivistic research commonly found in the literature.

After acknowledging the above facts regarding the nature of qualitative research, let us define it. Qualitative research is an umbrella term which encompasses enormous variety. Flick (2009) has defined qualitative research as research where the researcher is, "interested in analyzing the subjective meaning or social production of issues, events or practices by collecting non-standardized data and analyzing texts and images rather than numbers and statistics."

In fact, qualitative research is more diverse in nature than quantitative research. In qualitative research the structure of the design and data are not organized in advance, rather they develop as the empirical work proceeds. Qualitative research is also much more eclectic in using multiple strategies and methods as compared to quantitative research. Despite these diversities in methodological strategies, qualitative research has some recurrent features (Tesch 1990; Wolcott 1992). Miles and Huberman (1994) have provided a useful summary of such features of qualitative research. An attempt is being made to present this summary as under:

(i) Qualitative research is conducted through a prolonged contact with a 'field' or life situation. These situations are reflective of the everyday life of persons, groups, societies and organizations. In other words, it is *naturalistic* preferring to study people, things and events in their natural settings.

(ii) The researcher aims to gain a holistic view of the context under study, that is, its logic, arrangements as well as explicit and implicit rules that are all under the target of overview.

(iii) The researcher attempts to gather data on the perceptions of the local individuals from the 'inside' through the process of a deep attentiveness of empathic understating and of suspending any bias or preconceptions about the topics under study.

(iv) After going through various materials, the qualitative researcher may identify certain themes and expressions, which can be reviewed with informants but that must be maintained in their original forms throughout the study.

(v) The major task of the researcher is to explicate the ways the individuals in a particular settings come to understand and take action and otherwise manage their day-to-day situations.

(vi) The researcher may make many interpretations of the materials but some may be more compelling for theoretical reasons or on the grounds of internal consistency.

(vii) Qualitative researcher uses little standardized instruments at the outset. The researcher himself is the main measurement device of the research.

(viii) Most analysis is done with words. The various appropriate words are assembled, subclustered, and broken into semiotic segments. These words are frequently organized to permit the research to compare, contrast, analyze as well as bestow patterns upon them.

A QUALITATIVE RESEARCH MODEL: FIVE COMPONENTS

In qualitative research the activities of collecting and analyzing data, developing and modifying theory as well as identifying and dealing with various validity threats usually go on simultaneously, each influencing all the others. Maxwell (1998) has proposed a model of qualitative research in which five components for conducting qualitative research have been outlined. In fact, these five components can be phrased as basic questions that the researcher needs to answer in conducting a qualitative research. These components as phrased in terms of questions may be discussed as under:

1. What are the goals of qualitative research?

The researcher sets the goal or purpose of the study or the questions that are hopefully to be answered. According to Maxwell (1998), there are five specific purposes of qualitative research.

(i) To understand the way the subjects or participants in the study understand the behaviours or events

(ii) To understand the context or environment in which the participants exist

(iii) To identify some unexpected phenomena and influences and generate new grounded theories

(iv) To understand those processes that lead to outcomes

(v) To develop causal explanations in rare cases

2. What are the contexts in which the study is done?

The researcher has to frame his conceptual context, that is, what he thinks is going to explain the behaviours, events, etc. under study. Here the researcher needs to understand appropriate theory, prior research and also some personal experiences from which he may be able to draw something.

3. What does the researcher want to understand at the end of the study?

In qualitative research, the purpose and context need to be understood for the research questions to be formulated. Like all components of the study, the questions are open to change while the study is going on.

4. What will the researcher actually do?

In simple words, what approaches, strategies and methods will be used to collect data. This also includes the sampling procedures, that is, the question of who is being studied and how the individuals are selected for the study. In qualitative research, the researcher prepares a tentative plan for the study. Maxwell has emphasized upon four components of methods used in qualitative research: the *relationship* that the researcher establishes with participants; *what settings or individuals* and *what other sources of the information* that the researcher decides to use; how the researcher gathers the information; and what the researcher will do with information to make sense of it.

So far as the relationship established between the researcher and the participant is concerned, it can be *complex* and *changeable*. The role of the researcher can change from being an observer to being a participant observer as we find in some ethnographic studies. The second component is concerned with sampling technique. In qualitative research, generally *purposive sampling* is used since this provides the information desired. The basic aim here is that the sample is useful in answering the questions raised by the researcher. The third component is regarding gathering of data, that is, in fact a broad topic. The common methods of gathering data are observation, interviews and document analysis.

5. What alternative explanations may fit the findings?

Why should the findings of qualitative research be trusted? What threats are there to the validity of the research? Following Maxwell, there are two critical aspects of research validity for qualitative research: the bias of the researcher and the impact of the researcher's presence on the participants or research settings. The researcher bias means the presence of some preconceived notions and beliefs that may affect data collection. The reactivity of the participants to the presence of the researcher is also important. Since both these aspects of qualitative research are in a way part of the study, the researcher must explain and understand them rather than making an attempt to eliminate them by concluding that they are not possible.

These five components as spelled out in terms of questions, tend to obtain answers which may change during the study and therefore, they are not linear in relationship with each other. Thus these five components form an integrated and interacting whole where each component is closely tied to others, rather than being linked simply in linear cyclic sequence.

RELEVANCE OF QUALITATIVE RESEARCH

Why one should use qualitative research? What is the special need to use such an approach in current situation? Has quantitative research failed in its mission?

The answers to these questions have hinted towards the significance and relevance of qualitative research. It has been argued that qualitative research is of much relevance for the study of social relations probably due to the *pluralization of life worlds*, which refers to the diversification of ways of living in one society. As we know, today the traditional class models in the society such as working class, middle class and upper class are no longer considered adequate for describing modern societies as many local, subcultural, etc., differences have became equally important. In such pluralization of life worlds the key components are new obscurity (Habermas 1967), the growing individualization of ways of living (Beck 1992) and the dissolution of old social inequalities into new lifestyles and ways of living. In the days of this pluralization, qualitative research that aims basically to study the participants in their natural settings has proved to be of immense importance.

Apart from this, postmodernism has also argued that the era of big narratives and theories is gone. As we know, postmodernism is a social theory that criticizes modernism and its concepts of facts and science and takes the way facts and science are produced into account (Flick 2009). Today locally, temporally and situationally limited narratives are the need of the day. These days, social scientists are giving more emphasis upon what is actually taking place in the sphere of life chosen for the study. Rapid social changes and resulting diversification of life have forced social scientists to use new social contexts and perspectives. The days of classical *deductive methodologies* that require hypotheses formulation from various theoretical models and testing them against empirical evidences are gone now because they are failing due to the differentiation of objects. These scientists are *forced* to make use of *inductive methodologies*. Instead of starting from theories and testing them, they require some sensitizing concepts for studying varying social contexts for which qualitative research is considered important and much relevant (Bruner 1991).

QUALITATIVE RESEARCH: A BRIEF HISTORICAL INTRODUCTION

If we pay attention to the history of qualitative research, it is obvious that psychology and other disciplines of social sciences have a long history of using qualitative research. History bears testimony that in psychology, Wilhelm Wundt used methods of description and *verstehen* in his *folk psychology* along with his experimental research work. More or less at the same time, a debate in German sociology was started between a science, which showed inclination towards induction as well as case studies, and an empirical and statistical approach. In American sociological methods such as biographical methods, case studies and descriptive methods remained central till the 1940s. It was not until the 1960s that in American sociology the criticism of standardized and quantifying social research became important (Glaser & Strauss 1967) and this led to renaissance of qualitative research in social sciences as well as in psychology with some delay (Willig & Stainton-Rogers 2007). As a consequence of the said development in Germany and United States, qualitative research progressed in phases. It is also important to note that many important psychological insights including the work of Piaget and Vygotsky have roots largely in qualitative research.

In the United States of America, Denzin and Lincoln (2005) have identified seven phases of the development of qualitative research. The *first phase* is that of traditional period ranging from 1900 to 1945. This period is related to the research of Malinowski (1916) and the dominant Chicago School of Sociology. During this period, qualitative researchers were interested in the other (that is, something foreign or strange) and in its objective description and interpretation. The *second phase* was that of *modernist* phase ranging from 1945 to the 1970s. During this period, serious attempt was made to formalize qualitative research by publishing more and more textbooks. The spirit has been kept still alive by Glaser and Strauss (1967), Strauss and Corbin (1990) and by Miles and Huberman (1994). *The third phase* is that of *blurred genres* that continued until the mid-1980s. During this period, the researchers placed various theoretical models as well as the understandings of objects and methods side by side so that they can choose and compare various alternative models such as ethnomethodology, phenomenology, symbolic interactionism, semiotics or feminism (Guba 1990). The *fourth phase* was *crisis of representation,* which began since the mid-1980s. The crisis of representation discussion done in the field of artificial intelligence and ethnography had great impact upon qualitative research as a whole. This made the process of displaying knowledge and findings a substantial part of the research process. Consequently, qualitative research moved towards a continuous process of constructing versions of reality. Whatever people narrated in the interview does not necessarily correspond to whatever they had formulated at the moment the reported *event* happened. Likewise, the version given at the interview may not necessarily correspond to the version they would have given to a different researcher with a different research question. It meant that the researchers, attempting to interpret the interview and present it as a part of their findings, tend to produce a new version of the whole. The *fifth phase* started in the 1990s. During this phase, narratives replaced theories or theories were read as narratives. Here the emphasis shifted towards such theories and narratives that fit specific, local and historical problems. The *sixth phase* is marked by post-experimental writing as well as linking issues of qualitative research to some democratic policies (Flick 2009). The *seventh phase* covers roughly the period of 2000 to 2004 and is characterized by establishing qualitative research through various successful and new journals. Later, evidence-based practice emerged as the new criterion of relevance for social science and is counted as the *eighth phase* in the development of qualitative research as claimed by Denzin and Lincoln (2005).

So far as the development in Germany is concerned, Habermas (1967) had first recognized that a new tradition and discussion of research was developing in American sociology related to names like Cicourel, Goffman, etc. Right from this period, the model of research process put forward by Glaser and Strauss (1967) attracted the attention of researchers. Here, the basic aim was to do, so far as possible, more justice to the objects of research in quantitative research. Kleining (1982) had emphasized upon understanding the object of research as preliminary till the end of the research because object would come to its form only at the end. A discussion about a *naturalistic sociology* and *the principle of openness* also provided strong impetus for growth and development of qualitative research (Hoffmann-Riem 1980). At the end of the 1970s, a still more broader discussion started in Germany. This discussion exclusively dealt with interviews, how to apply and analyze them as well as with methodological questions, which stimulated extensive research. The main question of this period was to determine whether these developments should be perceived as fashion or something like a new beginning.

In Germany, two original methods were picked up at the beginning of the 1980s and these two methods were considered crucial for the development of qualitative research. These two methods were: *narrative interview* and *objective hermeneutics*. As we know, narrative interview is a specific form of interview based on one extensive narrative. Here instead of asking questions, the interviewer asks the persons or subjects to tell the story of their lives as a whole, without interrupting them by asking questions. Objective hermeneutics is a way of doing research by analyzing the texts for identifying its latent meaning. For example, transcript of a family interaction may be analyzed for identifying and elaborating an implicit conflict underlying the communication of the members involved in interaction. Both these methods stimulated extensive research practice in biographical research (Rosenthal 2004). In the middle of the 1980s, researchers also started showing concern for the problems of validity and generalizability of the obtained results. The unstructured nature of the data obtained by the use of these methods also required the use of computers in qualitative research (Richards & Richards 1998; Gibbs 2007).

So far as the research in the qualitative tradition in India is concerned, it is relatively a recent development and is still in the process of consolidation (Prakash 2011). Since qualitative research questions many of the well-established assumptions without offering the certitude enjoyed by the quantitative methods, it is not very popular and therefore, only recently psychologists have started using the diverse types of qualitative research. In Indian psychology, qualitative research has been practical through interview (Goyal 1997; Dhawan et al. 1999; Chopra et al. 1998), observation (Babu et al. 2000; Kar et al. 2000), narratives (Bharat & Aggleton 1999; Anand et al. 2001; Johri 2001; Puri 2002), ethnography (Priya 2002; Sibia & Raina 2001; Vagrecha & Asthana 2002); case study (Dosajh & Dosajh 1999; Mangla 1994; Vohra 2001), textual analysis (Kishore 2001; Sharma 1995), participating rural appraisal (Singhal 2003; Anand 1998; Pratheep 1994), conversational analysis and focus group (Natrajan & Thomas 2002; Rangnathan 2003) and grounded theory and biographical analysis (Nandy 1995; Wali 1987; Raina,1997; Chitraprabha & Bhargava 2001).

Thus we find that the development of qualitative research in various countries has gradually reached a very crucial stage where, on some points, it is being considered more important than ever before.

THEMES OF QUALITATIVE RESEARCH

There are different types of themes which characterize qualitative research. According to Patton (2002), there are twelve themes which inculcate qualitative research. These twelve themes are divided into three major categories: *Design strategies, Data collection fieldwork strategies* and *Analysis strategies*. A discussion follows.

1. *Design strategies:* All qualitative research uses some design strategies. These research designs may be grouped under three heads: *Emergent design flexibility, purposive sampling* and *naturalistic enquiry.*

In emergent design flexibility, as its name implies, the researcher avoids getting located into any rigid design and he may bring a change in his research work as the data are being collected and the samples are selected for their usefulness. This design flexibility is a staple of qualitative research that distinguishes it from quantitative research where the entire research design is spelled out before any data are created and also, no changes are made during the study. A qualitative researcher has some ideas about the research design but it is open to change according to the situation and as data are collected and analyzed. This allows the researcher to make correction in design for adjusting to new information and the situation.

In qualitative research, the researcher gives preference to *purposive sampling* in comparison to the other form of sampling. In quantitative research, purposive sampling is rarely used. A qualitative researcher prefers purposive sampling (or a particular sample of persons or documents) because such persons are 'information rich' and 'illuminative'. In other words, they offer useful and desired information regarding the phenomenon of interest.

Another design strategy is *naturalistic inquiry*, also known as *ethnographic studies*. Here the researchers study real-world situations as they unfold naturally. They make observation without any attempt to manipulate or control the environment. The researchers' strategies show openness to whatever emerges.

2. *Data collection and fieldwork strategies:* According to Patton (2002), there are four data collection and fieldwork strategies in any qualitative research: *qualitative data, personal experience and engagement, empathetic neutrality and mindfulness* and *dynamic systems.*

Qualitative data include inquiry, in-depth interviews that capture direct quotations about persons' experience, case studies, observations yielding detailed and thin observation and careful document review. The very nature of qualitative data makes qualitative research very powerful. In fact, the richness of the obtained data permit an in-depth understanding of what is being studied than could be derived from qualitative research.

The strategy of *personal experience and engagement* means that the researcher has direct contact with people, situation and phenomenon under study by going into the real world of programme, organizations, neighbourhood and street corners. In fact, such direct contact makes possible description and understanding of both observable behaviours and internal states. Here the researcher's personal experiences and insights are considered an important part of the inquiry for understanding the behaviour under study.

Empathic neutrality and mindfulness is another important strategy. Here the researcher has to be empathetic and neutral at the same time in dealing with his participants or subjects. Of course, this is a difficult task. The researcher tries to understand the feelings of the subjects as well as show empathy towards them by displaying openness, sensitivity, respect and responsiveness while at the same time maintaining perspective as a good researcher. The researcher tends to report on any sources of bias and uses multiple sources for producing trustworthy results.

Dynamic system assumes that changes are an on-going process whether focus is on the individual or an entire community. A single-time picture of the individual or community as a whole is not enough and often prove to be misleading. Any individual or community can be understood only by observing the various changes that occur over the course of the study. The qualitative researchers observe these changes over time and find a link for the various changes and processes that occur.

3. *Analysis strategies:* There are some analysis strategies in qualitative research. These are *unique case orientation, inductive analysis, holistic perspective, context sensitivity* and *voice, perspective and reflexivity.*

Unique case orientation means that the researcher treats each case as a unique one. The researcher's own training and background regarding the case also adds to the unique quality of the data. For example, the data from a case study of a mentally retarded infant, and therefore the resulting report, would differ on whether the researcher's background and training were in the field of medicine or social work or special education. As a result, what the reader will know about the case would obviously depend upon the data collected and the perspective of the researcher.

Inductive analysis is another analysis strategy that allows the researcher to explore and analyze the data without prior hypotheses. It helps the researcher to discover reality without fitting it into a preconceived theoretical perspective. Such openness to find whatever there is to find has been considered unique to qualitative research and makes it distinct from quantitative research, which is based on hypotheses generated from theory, prior research or some models.

Holistic perspective is another strategy. Here the researcher concentrates on the whole phenomenon under study that is considered as more than a sum of its parts. Here the focus is on complex interdependencies and system dynamics that cannot be reduced to a few discrete variables and cause-effect relationship. Thus, this perspective is important for understanding the complex nature of many aspects of human behaviour.

Context sensitivity is another analysis strategy. It means that qualitative data are very sensitive to the social, historical and temporal context in which they were collected. That is the reason why data are not generalized to other contexts socially, spatially or temporally. Since something occurs in one family at a particular time does not mean that a similar thing should occur in some other family also.

Voice, perspective and reflexivity is other important strategy. A qualitative researcher needs to be reflective about his own voice and perspective. Here reflexivity is used in a wider perspective, which requires the qualitative researcher to be attentive to the cultural, political, social, linguistic and ideological aspect of his own perspective and voices of those interviewed and to those for whom the report is prepared.

Thus we find that themes of qualitative research are broader ones and this also show their distinctions from quantitative research on several points.

THEORETICAL PERSPECTIVES OF QUALITATIVE RESEARCH

There are different theoretical perspectives of the methodologies subsumed under the term qualitative research. According to Jacob (1987) there are six such perspectives: two from ecological psychology (behavioural settings and specimen records), three from anthropology (holistic ethnography, ethnography of communication and cognitive anthropology) and one from social psychology (symbolic interactionism). Although Jacob's perspective system is useful, it is not comprehensive. At the same time, it complicates the issue by fragementing some fields like two types of ethnography and two areas of ecological psychology.

A much more wider system of categorizing theoretical perspective has been provided by Patton (2002). He has provided 16 theoretical perspectives of qualitative research. A brief discussion follows.

1. *Ethnography*: The disciplinary root of ethnography is anthropology, which includes the study of cultural phenomena. Ethnography, besides being a theoretical perspective, is also a method of doing qualitative research. The major purpose of ethnographic research is to determine the physical and social environment in which the persons under study, live and work. As a research strategy, the researcher here combines different methods but bases it on participation, observation and writing about the field or situation under study. For example, the qualitative researcher wants to study how homeless and parentless adolescents deal with their health problems. Here he would combine participant observation in their community with their interviewing. The resulting picture of the details from the participation, observation and interviewing is unfolded in a written text about the field.

2. *Autoethnography:* Autoethnography stems from literary arts. Here the researchers are interested in answering questions like: how does own experience of this culture connect with and tend to offer insights into this culture, event, situation and the way of life of the inhabitants?

3. *Reality testing:* Positivist and Realist approach—Reality testing stems from philosophy, social sciences and evaluation. Here the researcher is interested in finding answers to several questions like: what is really going on in the real world? What is the truth in so far as we can get it? How can we study a phenomenon so that the conclusion may correspond to the real world as far as possible?

4. *Constructionism/Constructivism:* The disciplinary root of constructionism/constructivism is sociology. Here the researcher's aim is to answer the question: how have the individuals in a setting or situation constructed reality? What are their reported perceptions, beliefs and explanations? What are the consequences of their constructions for their behaviour?, etc.

5. *Phenomenology:* Phenomenology stems from philosophy. Here careful description and analysis of the individual's life world, and the meaning making and understanding in that life world are done. The researcher is basically interested in answering questions like: What is the meaning, structure and essence of the lived experience of a phenomenon for the persons or group of persons?

6. *Heuristic inquiry:* Heuristic inquiry stems from humanistic psychology. Here the researcher tries to answer the question: What is the experience of others who have also experienced the same phenomenon intensely?

7. *Ethnomethodology:* Ethnomethodology stems from sociology. Here the major purpose of the researcher is to answer the question: In what ways people make sense of their everyday activities so as they may be able to behave in socially acceptable ways?

8. *Symbolic Interaction:* Symbolic interaction stems from social psychology and is also used in linguistics to the same extent. This theory in qualitative research is based upon the assumption that people act and interact on the basis of meaning of words, things and their interpretation. Another assumption is that the meaning of objects, things, etc., is derived from the social interaction that a person has with others. Still another assumption has been that these meanings are modified through the interpretative process used by the persons in dealing with things, objects, etc., they encounter (Blumer 1969). Here the researcher tries to answer the question: What common set of symbols and understandings has emerged to give meanings to the interactions of the individuals?

9. *Ecological psychology*: Ecological psychology emerges from ecology and psychology. Here the researcher tries to answer a central question like: How do persons try to accomplish their goals through specific behaviours in specific environments? This perspective attempts to understand the connection between persons' behaviour and their environment. Ecological psychologists strongly emphasize that behaviour is goal-directed. They collect data by observing the person and the environment. They try to keep detailed descriptions of both and then code these descriptions for numerical analysis. In this way, methodology is qualitative whereas the data analysis remains quantitative.

10. *Semiotics:* Semiotics stems from linguistics. Here the basic question to consider is: How do various signs such as words, symbols, etc., convey meaning in a particular context?

11. *Hermeneutics:* The disciplinary roots of hermeneutics are linguistcs, philosophy, literary criticism and theology. Here the central question is: What are the conditions under which a human act takes place that helps in interpreting its meaning?

12. *Narrative analysis/Narratology:* Narrative analysis/narratology stems from social sciences, literary criticism and literary nonfiction. Here the central questions before the researchers are: What does the story or narrative reveal about the individual and the environment from which it came? How can one interpret this narrative for understanding and illuminating the life and culture that created it?

13. *System theory:* System theory stems from interdisciplinary approach. The question with which the researcher is concerned here is: How and why does this system as a whole function as it does?

14. *Chaos theory:* Nonlinear dynamics—Chaos theory stems from theoretical physics and natural sciences. Here the central question of concern is: What is the underlying order, if any, of disorderly phenomenon?

15. *Grounded theory:* Grounded theory stems from social sciences and methodology. Grounded theory is a method, an approach, a strategy and therefore, it is not a theory at all. It can best be defined as a research strategy whose basic purpose is to generate theory from data. 'Grounded' clearly means that the theory will be generated on the basis of data. Theory will, therefore, be grounded in data. Theory here means that the objective of collecting and analyzing data obtained from research is to generate theory. Therefore, the essential data in grounded theory is that theory would be developed inductively from data. Grounded theory, besides being a theoretical perspective, is an overall strategy for doing research. Here the central question of concern is: What theory actually emerge from systematic comparative analysis of the obtained data?

16. *Orientational:* Feminist inquiry, Critical theory, etc.—Feminist inquiry, critical theory, queer theory, etc., stem from political, cultural and economic ideologies. Here the central question of concern is: How feminist perspective manifest in the phenomenon under study? As

we know feminist thinking has been making an important contribution to social science research since 1960s, playing a particularly very significant role in qualitative research. There are different feminist approaches to research methods (Olesen 1994; Roman 1992). Feminist research focused on ignorance of women's life situation and of male dominance. Such research was often marked by the use of qualitative research due to the methods opening up more to women's voices and needs in general. Mies (1983) has pointed out reasons why feminist research is more linked to qualitative research rather than to quantitative research. Quantitative research often ignores the voices of women, turns them into objects and theory are often studied in a value-neutral way. Qualitative research allows women's voices to be heard and goals realized. As Ussher (1999) has pointed out, feminist research focused on a "Critical analysis of gender relationships in research and theory... and the recognition of need for social change to improve the lives of women."

Thus we find that there are different theoretical perspectives. Of these various perspectives, three theoretical perspectives namely ethnography, symbolic interactionism and ecological psychology are very common ones.

RESEARCH DESIGN STRATEGIES OF QUALITATIVE RESEARCH

As we know, research design is a plan for collecting and analyzing evidences, which enables the researcher to answer whatever questions he has posed. In fact, research design aims to touch almost all aspects of the research, from the minute details of data collection to the selection of the various techniques of data collection. A qualitative research generally uses the following popular research design strategies:

1. Ethnography or ethnographic studies
2. Case studies
3. Document or content analysis
4. Grounded theory
5. Retrospective studies

1. *Ethnography or ethnographic studies:* Ethnography*, sometimes known as *cultural anthropology* or recently as *naturalistic inquiry* is, in fact, a method of field study observation that became popular in the second half of the 19th century. The term 'ethnography' itself comes from cultural anthropology. 'Ethno' means people or folk whereas 'graphy' means describing something. Thus ethnography means describing a culture and understanding a way of life from the viewpoint of its participants. In still very simple words, it can be said that ethnography is the study of people in naturally occurring setting or fields by means of methods that capture their ordinary activities and their social meanings involving the investigator participating directly in the setting for collecting data in a significant manner without anything imposed on them externally.

Hammersley and Atkinson (1995), prominent authorities in the field have described ethnography by taking a fairly liberal view, whereby the researcher participates, overtly or covertly, in the people's daily lives for an extended period of time, watching what happens, listening to what is said, asking questions and collecting relevant data. They have emphasized ethnographic connection to naturalism, a way of doing social research developed by ethnographers in face of difficulties created by positivism.

Some important studies involving ethnography have been done in the Indian context. Sarangpani (2003) conducted an ethnographic study of teacher-taught relationship in a village in North Delhi. It was an interaction between those who know (teachers) and those who came to know (schoolchildren). Children perceived schooling as something that makes them good citizens and empowers them to move away from traditional jobs. Teachers' voice and memorization were considered important. Memorization and recalling were perceived as the key criteria of intelligence. Priya (2002) conducted an ethnographic study for investigating the processes of suffering and healing among the victims of Bhuj earthquake that brought havoc on 20 January 2001. For understanding the process of suffering and healing, she used a semi-structured interview, focus group and observation. It was reported that the villagers expressed a strong belief that doing their duties (karmas) would bring harmony and peace, which would lead to healing among survivors. On the other hand, victims in urban areas were, in fact, more materialistically-oriented. Likewise, in an ethnographic study of educational institution named Mirambika at Delhi, Sibia and Raina (2001) utilized participant observation, interviews, informal interactions and questionnaires over a period of eight months. Thematic data analysis revealed that learning features and interpretations of classroom were drawn by inductive analysis. Validity of the interpretation was established through triangulation. The study provided a kaleidoscopic view of curriculum organization, teaching styles and disciplinary techniques and seating patterns of classrooms of Mirambika. Similarly, using ethnographic field study in Bhubaneswar, the domestic lives of women and families have been investigated by Menon and Shweder (1998) and Seymour (1999). In these studies a fine blending of observation, interview and narratives was done and they offered an insightful mapping of the self and family situated in the social context.

There are some features of ethnographic studies. The central feature of ethnography is to study and understand the cultural and symbolic aspects of behaviour and the context of that behaviour, whatever may be the specific focus of research. In ethnographic study the specific focus of research is typically either some group of people or a small number of cases or just one case, in detail. In addition to these central characteristics, ethnographic studies have the following six characteristics.

(i) The ethnographer starts with the assumption that the shared cultural meanings of group are crucial to understanding its behaviour. Any group of persons like prisoners, tribals, patients, etc., develop lives of their own that become somehow meaningful and reasonable once the researcher get close to it. The ethnographer's primary task is to uncover such meaning (Goffman 1961).

(ii) In ethnographic study the researcher is sensitive to the meanings that behaviour, actions, events, etc., have in the eyes of the people being studied. Therefore, a major task of the ethnographer is to elicit that meaning or knowledge from insider's perspective.

(iii) The ethnographer studies the group or case in its natural settings. True ethnography, therefore, involves the researcher becoming part of that natural setting (Fielding 1996)

(iv) Ethnography is not a prestructured study, rather an unfolding and evolving part of study. It means that it will not be normally clear what to study in depth until some progress in research work has been done. While specific research problems and hypotheses are used in the research, they are more likely to develop as the study proceeds. This principle also applies to data-collection procedures and to data.

(v) Ethnographic data collection is typically *prolonged* and *repetitive*. There is both specific and general reason for this (Woods 1992). The specific reason is that the ethnographer needs a comprehensive and detailed record and typically focuses on things that happen time and again. The general reason is that since reality is being studied, the meanings, symbolic importance and cultural interpretation exists on several levels. Naturally, it takes time for the ethnographer to arrive at the deeper and important levels of reality.

(vi) From the viewpoint of data collection, ethnography is eclectic and not restrictive. The continuum of ethnographic field work may range from direct nonparticipant observation to participant observation, then to interviewing with one or more informants, and then to the words of people themselves.

---

* For a further discussion, the reader is referred to consult chapter 15.

When is ethnographic approach considered most appropriate? In general, an ethnographic approach is considered most appropriate when we need to understand the cultural context of behaviour and the symbolic meanings and the significance of the behaviour within the context. In fact, it is an excellent way of developing insight into a culture or social process, particularly those involving other cultures and subcultures including those of organizations and institutions.

The major limitation of ethnographic study is the danger of methodological arbitrariness despite the fact that the approach shows flexibility towards the subject under study. Many early ethnographic studies were criticized on the grounds that the ethnographer spent too little time among the people of the tribe or individuals to be studied to get more than superficial view. Later, investigators developed clear realization that ethnographic studies would be meaningless and invalid if the researchers have not met the following conditions:

(i) The researcher must live for a longer and extensive period of time among the individuals/tribe and become an integrated member of the cultural group.

(ii) The researcher must learn the native language of the cultural or social group so that he should be able to develop the sensitivity to think, feel and interpret observations in terms of the cultural group's concepts, feelings and values.

(iii) The researcher must have trained his informants to systematically record field data in their own language and cultural perspective.

2. *Case studies:** In qualitative research, like quantitative one, case studies are the popular research strategies. A case study is a way of organizing social data for the purpose of knowing about social reality. It examines social unit as a whole and the unit may be a person, a family, a community, a social group, a social institution or a nation. It could also be a decision, or a policy or a process or an event or incident of some sort and there are also other possibilities as well. According to Brewer and Hunter (1989) there can be six types of units (or cases) which can be studied: individuals, attributes of individuals, action and interactions, residues and artifact of behaviour, settings, incidents, and events and collectivities. Any of these units can be in focus in a case study.

Obviously, then, there are different types of cases and therefore, there can be different types of case studies. Stake (1994) has distinguished between the following three types of case studies:

(i) *Intrinsic case study:* It is defined as a case study where the study is undertaken because the researcher prefers a better understanding of the case selected for study.

(ii) *Instrumental case study:* It refers to a case study where a particular case is examined to give insight into an issue, or to refine a theory.

(iii) *Collective case study:* It refers to a case study where the instrumental case study is extended to cover several cases as well as to learn more about the general condition. It is also called as *multiple case study* or *comparative case study.*

The first two types of case studies are single case studies where the researcher gives focus within the case. The third type of case study involves multiple case study where the researcher gives focus upon both within and across cases.

Various definitions of case study have been offered. According to Stake (1988), case study is defined as a study of bounded system, emphasizing the unity and wholeness of that system but paying attention only to those aspects that are relevant to the research problem. Yin (1984) has defined case study as an empirical inquiry that investigates a contemporary phenomenon within

its real-life context when the boundary between the context and phenomenon is vague and not clearly evident and in which multiple sources of evidence are used. Following Goode and Hatt (1981), case study is not a specific technique rather it is a way of organizing social data so as to preserve the unitary character of the social object to be studied.

Analyzing these various definitions, we get the following *four* characteristics of case study :

(i) A case is a *bounded system,* that is, it has boundaries. Generally, the boundaries between the case and the context are not necessarily evident but the researcher needs to identify and describe the boundaries of the case as evidently as possible.

(ii) The case is case of something. The researcher needs to identify what the case is, a case of importance in determining the unit of analysis.

(iii) In case study there is an obvious attempt to preserve the wholeness, unity and integrity of the case.

(iv) In case study, multiple sources of data and multiple data collection methods are used in naturalistic setting. Some case studies prefer to use sociological and anthropological field methods such as observations in natural settings, interviews, and several narrative reports whereas some case studies may use questionnaires and some kind of numerical data. It obviously means, then, that case study is not necessarily a qualitative technique, it is also used as a quantitative technique. However, most case studies are predominantly qualitative.

Yin (1998) has listed six types of data with which researchers of case study must be aware.

(a) *Archival record:* Archival records are usually quantitative data that include survey, questionnaires, records of organizations, results from prior studies, etc.

(b) *Documents:* Documents include materials like school records, health records, official notes, etc.

(c) *Direct observation:* When the researcher visits the case study site, he prepares some field notes based upon his observation.

(d) *Participant observation:* It is observation of the field situation by the researcher who becomes a participant in the situation being observed. This provides important data for some case studies such as when a college student wants to know how does a seminar meant for college students operates.

(e) *Interview:* In case studies, interviews are open-ended in nature. The interviewers are considered as informants for these data and the verbal reports of the interviews should be triangulated for later verification.

(f) *Physical artifacts:* This is used in physical anthropological studies of earlier times. Arranging a classroom to understand the class as a case would be an example.

Those willing to conduct research using case study need to follow the steps as described below.

(i) They should be clear on what the case is, including identification of its boundaries.

(ii) They should be clear on the need for the study of this case as well as regarding general purpose of the concerned case study.

(iii) They should translate general purpose into specific purpose and research questions.

(iv) They should identify the overall strategy of the case study, especially whether it is one case or multiple case.

(v) They should clearly specify *what* data are to be collected and from *whom* and *how.*

(vi) They should make explicit how the data will be analysed.

Indian Psychologists have used case studies in different settings such as in clinical setting, organization, etc. Dosajh and Dosajh (1999) used case study to examine a case of 45-year-old man with two children who worked as assistant manager in a bank. He was on medical leave and was planning to take premature retirement due to intense pain in the lumbar region. It was found that his disturbed family life, deprived sexual relations and depression had played a significant role in his illness. As a therapeutic measure, the senior officers were requested to change their attitudes towards the patient. The patient recovered and resumed his duties after proper cooperation rendered by colleagues and members of the family. Likewise, case studies have been used by some researchers to study women's problems such as physical assault (Gandhi 1995) as well as the effectiveness of compromise on the complainant (woman victim) and the offenders (such as husband and in-laws) (Chopra 1995).

---

* For a further discussion, the reader is also referred to consult chapter 15.

3. *Document or content analysis:* The reader is referred to consult chapter 12 for a detailed discussion of document or content analysis.

4. *Grounded theory:* Grounded theory is both a research strategy and a way of analyzing data. In this section, we shall discuss grounded theory only as a research strategy and in the next section, we shall deal with it as a way of analyzing data. As it has been explained earlier, grounded theory was developed by Glaser and Strauss (1967) when they were doing research from a sociological perspective in organizational contexts. In other words, it was developed as a method for the study of social behaviour, especially, complex one. The essential idea in grounded theory is that theory will be developed inductively from data. Grounded theory, thus, is an overall strategy for doing research. For implementing that strategy, grounded theory has a particular set of techniques and procedures.

Now let us look at the historical background of grounded theory. The history of grounded theory is a short one and can be traced mainly through its five key publications. In 1960, Glaser and Strauss started collaborative work in medical sociology and published two very important studies of dying in hospitals (Glaser & Strauss 1965; 1968). These two publications were *Awareness of Dying* in 1965 and *Time for dying* in 1968. These two books proved to be very important ones and represented a different style of empirically based sociology. In response to the numerous requests made by the readers of those two books, Glaser and Strauss wrote another book and it was published in 1967 under the title *The Discovery of Grounded theory*, which was regarded as the first important publication about grounded theory. This book had three major purposes to serve: to offer the rationale for theory that was grounded, to suggest the logic for and some specifics of grounded theory and to legitimate qualitative research carefully. Later on, Glaser and Strauss taught grounded-theory-style seminar in qualitative analysis at the University of California. Many researches using grounded-theory-to investigate a variety of phenomena were published later. The second key publication entitled *Theoretical sensitivity* was released by Glaser in 1978, almost 11 years after the publication of *The Discovery of Grounded theory*. The book focused mainly upon updating methodological development in grounded theory and to help the researcher develop theoretical sensitivity. The third key publication entitled *Qualitative Analysis for Social Scientists* came out in 1987, nine years after the second key publication. In this book, focus was given upon qualitative analysis in general. This book was regarded as the best handbook for providing better understanding of social behaviour through a particular style of qualitative analysis of data.

The fourth key publication entitled *Basics of Qualitative Research* and subtitled as *Grounded Theory Procedures and Techniques* was done by Strauss and Corbin in 1990. This book proved very useful to all those researchers in different disciplines who aimed to build theory on the basis of qualitative data analysis. In fact, the book presented the analytic mode of grounded theory. The fifth important publication was the result of revision of Strauss and Corbin's book by Glaser. The book written by Glaser was titled as *Basics of Grounded theory Analysis* and subtitled as *Emergences Versus Forcing*. This book was published in 1992. In this book, Glaser has tried to correct all those misconceptions about grounded theory that were evident in Strauss and Corbin's book.

Besides these methodological works on grounded theory, others having impact upon development of grounded theory could also be mentioned like those by Charmaz (1983) as well as by Chenitz and Swanson (1986). Mention should also be made of two other works by Glaser—*Examples of Grounded theory: A Reader* (1993) and *More Grounded theory Methodology: A Reader* (1994) through which his criticism of Strauss and Corbin work continued. Finally, Strauss and Corbin presented an overview of grounded theory methodology in a chapter of book entitled *Handbook of Qualitative Research* edited by Denzin and Lincolm in 1994.

Although grounded theory initially presented as a research strategy and method of analysis of qualitative data, it is equally applicable to quantitative data in Glaser's view (1992). Originally developed in sociology, grounded theory is, in fact, a general way of approaching research and does not depend on any particular disciplinary perspective.

Grounded theory has its explicit purpose of the generation of theory from data. In grounded theory that aims to generate theory, no up-front theory is proposed nor any hypothesis is formulated for testing. In fact, the researchers have to start with an open mind aiming to end up with a theory.

So far as the use of literature in grounded theory is concerned, it has different perspective in comparison to other research approaches. The difference lies in the fact that how the literature is dealt with and when it is introduced and follows the principle that grounded theory places on theory generation. If a satisfactory theory already exists on a particular topic, there is no sense in undertaking a study to generate a new theory about it. The obvious rationale for doing a grounded theory study is that the researcher has no satisfactory theory on the topic. In such a case, the researcher wants to approach the data as open-mindedly as possible, guided by research questions. The problem with reviewing the literature in advance of such a study is that it may strongly influence the researcher when he actually begins working with the data. In such a case, it makes sense in delaying the review of literature at least until the conceptual directions within data have become clear. Here the researcher will introduce the literature later, seeing the literature as further data for study. Therefore, the key principle in using review of literature in grounded theory is that literature is seen as further data to be fed into analysis but at a stage in the data analysis when theoretical directions have become more or less clear.

5. *Retrospective studies:* Retrospective studies, as their name implies, are defined as those studies in which retrospectively from the point in time when the research is carried out, some events and processes are analyzed in respect of their meanings for different persons or for their collective life histories (Flick 2009). Biographical research is a good example of retrospective studies. Here those persons are selected for informants who will be meaningful and relevant for the process to be investigated. Data are often (but not necessarily) collected with the help of narrative methods. They are analyzed with narrative or hermeneutic approaches. The basic aim of the research here is to develop theories from data that are analyzed.

One danger in any retrospective research is that the current situation in which an event is recounted, tends to overlap the earlier situation or influences the assessment of the past event. Moreover, the perspective on the processes that are analyzed, may be distilled from the viewpoint of interviewers or from studying such documents, which have been produced or filed.

Of these various research design strategies, the qualitative researcher may select any one or may combine any two (or sometimes more than two) designs depending upon the research problem being investigated.

SAMPLING TECHNIQUES OF QUALITATIVE RESEARCH

In qualitative research sampling is a very important step, as it is in quantitative research. However, there is difference in sampling of these two approaches. In quantitative research, the focus is on *people sampling*. Here most often, probability sampling is used that is directed at representativeness. Because of this representativeness, the findings from the sample are generalized to the population. Qualitative research rarely uses probability sampling but rather uses *deliberate sampling*—purposive sampling is the term frequently used. It obviously means that sampling is done in a deliberate way, with some purpose in mind. This technique selects certain persons, settings or events because they can readily provide desired information. The basic aim is that the sample is useful in answering the questions raised by the researcher.

There is no simple typology of strategies for sampling in qualitative research due to the great variety of research approaches, purposes and settings. Miles and Huberman (1994) have listed 16 qualitative sampling strategies in typology. Patton (1980), Johnson (1990) and Janesick (1994) have contributed still other strategies. In general, the following 12 sampling strategies are very popular in qualitative researches.

(i) *Complete collection:* In this sampling strategy, the structure of group taken into account is defined before data collection. Gerhardt (1986) used this sampling strategy in the study about events and courses of patients' careers in chronic renal failure in south-east of Britain. In order to know about events and courses of such study, she decided to do a complete collection of all patients (male, married, 30 to 50 years at the beginning of the treatment) of the five major hospitals (that is, renal units) serving the south-east of Britain. Here we find that sampling is limited in advance by certain criteria: a specific disease, a specific age, a limited period, a specific region and a particular marital status. On the basis of these criteria, relevant cases are taken. Here sampling is carried out because virtual cases, which don't meet one or more of these criteria, are excluded in advance. Such method of sampling is common in rural studies.

Since in this sampling strategy the structure of the groups taken into account is defined before data collection is done, it restricts the range of variation in possible comparison. Therefore, at least at this level, there will be no real new findings.

(ii) *Theoretical sampling:* Theoretical sampling developed by Glaser and Strauss (1967) in medical sociology, is a sampling procedure in grounded theory research where cases, groups or materials are taken according to their relevance for the theory that is developed and on the background of what is already the state of knowledge after collecting and analyzing a certain number of cases. Paraphrasing the viewpoints of Glaser and Strauss regarding theoretical sampling, it can be said that it is basically a process of data collection for generating theory whereby the researcher jointly collects, codes and analyzes his data and decides what type of data be collected next and where to find them in order to develop his theory as it emerges. Such a process of data collection is controlled by what is called as emerging theory. The most important principle of theoretical sampling is the idea that subsequent data collection should be guided by theoretical developments that emerge in the analysis. In other words, the basic principle of theoretical sampling is the genuine and typical form of selecting cases or materials in qualitative research.

Sampling decisions in theoretical sampling may start from the level of groups to be compared or they may directly focus on specific persons. A very important question in theoretical sampling is how to decide when to stop integrating further cases. Glaser and Strauss have suggested the criteria of *theoretical saturation* (of a category, etc). By theoretical saturation is meant a point in the grounded theory research at which more data about a theoretical category does not produce any further theoretical insight (that is, nothing new emerges). Sampling and integrating further material is finished when a theoretical saturation of a category or group of cases has been reached.

(iii) *Extreme case sampling:* This sampling technique is used when the strategy is to learn from unusual or negative manifestations of the phenomenon. Here the field under study is disclosed from its extremities to arrive at an understanding of the field. Thus here, the researcher purposively integrates extreme or deviant cases. In a nutshell, the researcher here seeks cases that differ from the dominant pattern or that differ from the predominant characteristics of other cases. Here, the goal is to locate a collection of unusual, different or peculiar cases having special characteristics. That is why extreme case sampling is also known as *deviant case sampling*. Since cases are selected because they are unusual, the researcher hopes to learn more about the real social life of these cases. For example, suppose the researcher is interested in studying high-school dropouts. Let us suppose that the previous research suggests that majority of such dropouts come from families having low income, single parent and from those families that are geographically mobile. A researcher using extreme case sampling would seek majority-group dropouts who are from stable two-parent, upper-middle income families who are geographically stable.

(iv) *Typical case sampling:* Here the researcher selects some typical cases, which may represent the average or majority of the cases. Here, the field is in a way disclosed from center and from inside. For example, for studying about the success or failure of the foreign policy of the country, the researcher may select a few parliamentarians who are likely to represent the overall and average view of the public.

(v) *Maximum variation sampling:* A maximum variation sampling is one where the researcher deliberately seeks as much variation as possible in the samples. Thus here, the researcher integrates only a few cases but these are as different as possible so that the range of variation and differentiation in the field may be disclosed.

(vi) *Intensity sampling:* Here the researcher plans to select cases, events, etc., according to the intensity of features, processes and experiences, and so on as given or assumed in them. The researcher prefers cases with the greatest intensity or different intensities which are systematically integrated for the purpose.

(vii) *Sensitive case sampling:* Sometimes the researcher may select important or sensitive cases only in the sample for the desired outcome. For example, the researcher may select and integrate a few top leaders from Hindu and Muslim community to know whether the construction of a mosque at the designated place would be in the best interest of the public.

(viii) *Critical cases sampling:* Here sampling plan includes selection of such critical cases in which the relations to be studied becomes clear in the opinion of the experts or which are particularly important for the functioning of the programme to be evaluated.

(ix) *Convenience sampling:* Here the researcher selects those cases that are easiest to access under conditions of study. Such samples are cheap and quick. The person-on-the-road interview conducted by television programmes is an example of convenience sampling (or sometimes known as *haphazard sampling*). Television interviewers go on the road with camera and microphone to talk to a few people who are conveniently available.

Morse (1998) has also provided some useful suggestions regarding sampling. Her suggestion is clearly aimed at making a distinction between *primary selection* and *secondary selection*. She has outlined several general criteria for qualifying a good 'informant'. These criteria are relevant for selecting meaningful cases. This criteria included that the selected cases should have necessary knowledge and experience of the issue; they should have the capability to reflect and articulate, they should have adequate time to be observed and should be ready to participate in the study. Integrating cases according to these criteria is known as *primary selection*, which is different from *secondary selection*. By secondary selection is meant all those cases that somehow don't meet all the criteria previously mentioned but are willing to give some time for interview. Morse has cautioned that one should not invest many resources in such cases.

Experts have also provided some questions against which to check a qualitative sampling plan. For example, Miles and Huberman (1994) have suggested six general questions against which the effectiveness of any sampling plan can be judged. There questions are:

(i) Is the sampling relevant for the conceptual framework and research questions?
(ii) Will the phenomena, in which the researcher is interested, appear in principle?
(iii) Does the sampling plan enhance the generalizability of the findings?
(iv) Is the sampling plan feasible in terms of time, money and access to people?
(v) Is the sampling plan ethical in terms of such issues as informed consent, potential benefits and risks and the relationship with informants?
(vi) Can believable descriptions and explanations be produced which are true to real life?

DATA COLLECTION TECHNIQUES IN QUALITATIVE RESEARCH

Qualitative researchers generally study spoken and written representations and record of human experiences, using multiple methods and multiple sources of data. There are several techniques of data collections and several types of data collection might well be used in a qualitative project. Here we shall concentrate upon the following three most popular data collection techniques of qualitative research.

1. Interview
2. Observation
3. Document review

1. Interview

Interview is one of the main data collection techniques in qualitative research. Interview is used to gather information regarding the interviewee's experience and knowledge, his or her opinions, beliefs and feelings as well as demographic data. Interview questions can be asked so as to determine past or current information as well as predictions for the future. While interviewing is basically about asking questions and receiving answers, there is much more to it than that especially in qualitative research. Following Fontana and Frey (1994), interviewing has a wide variety of forms and multiplicity of uses. The most common form of interviewing is individual and face-to-face exchange but it may take the form of face-to-face group interviewing, self-administered or mailed questionnaire, telephone survey, etc. Interviewing can be structured, semi-structured or unstructured. An interview can be one-time brief exchange (say for 5 to 10 minutes) over telephone, or can take place over multiple, lengthy sessions over days as it happens in life-history interviewing.

There are different types of interviews and much has been written on this topic. For example, Patton (1980) has distinguished between three main types of interviews: *informal conversational interview, general interview guide* and *standardized open-ended interview*. Fielding (1996) described the types of interviews using the terms *standardized, semi-standardized* and *nonstandardized*. Similarly, Fontana and Frey (1994) used a three-way classification of structured, semi-structured and unstructured interview and they apply to both individual and group interview. Besides these, there are many other classifications of interview. Here we will discuss some of the common types of interviews frequently used in qualitative research.

(i) *Structured Interview:* A structured interview is one where the respondent is asked a series of pre-established questions with preset response categories. The series of questions asked in the interview is usually called as interview schedule. If the interview is standardized as well, all respondents receive the same questions in the same order. Therefore, flexibility and variations are minimized. Here the interviewer plays a neutral role. The stimulus response nature of this type of interview stresses rational rather than emotional information, which is easily quantified, ensures comparison of questions across respondents and makes sure that certain topics have been included. However, such an interview leaves little room for unanticipated discoveries. Sometimes the respondents feel constrained because they are not free to give the information that they feel is important.

(ii) *Unstructured interview:* In unstructured interview, the precise questions and their order are not fixed. They are allowed to develop as a result of conversation with the respondent. Open-ended answers allow interviewees to say as little or as much as they choose. This is also known as *informal conversation interview*. The traditional type of unstructured interview is the nonstandardized, open-ended, in-depth interview, sometimes also called as *ethnographic interview*. According to Fontana and Frey, there are seven important aspects of unstructured interview. These seven aspects provide a useful checklist of things to be done when planning the data collection by means of unstructured interview. The seven aspects are: *accessing the setting, understanding the language and culture of respondents, deciding on how to present oneself, locating an informant, gaining trust, establishing good rapport* and *collecting empirical materials*.

How each aspect is handled varies with the nature of the situation and respondents. There needs to be flexibility in unstructured interview situation, especially for life-history projects. Douglas (1985) has called this *creative interviewing*.

Analysis of unstructured interview is time consuming and difficult. Skill in this sort of interviewing does not come naturally and most of us need specific training to develop that skill. Here different types of information are collected from different people with different questions. This makes data organization very difficult.

(iii) *Group interview:* In group interview, the researcher works with several people simultaneously. Such an interview is also known as *focus group interview*. Originally focus group interview was a particular type of group interview used in marketing and political research. Group interview can make very important contribution in social science researches. According to Morgan (1988), the hallmark of focus group interview is the explicit use of the group interaction to produce such data and insight that would be less accessible without interaction found in a group.

There are different types of group interviews and like other types of interviews, they can be unstructured, semi-structured or highly structured. Different types of group interviews have different purposes. Which type of group interview should be used in a particular research situation depends on the context and research purposes.

The role of the interviewer or researcher changes in a group interview, functioning more as a moderator or facilitator and less as an interviewer. Here the process will not be putting question and answer in alternate way rather the researcher will be facilitating, moderating, monitoring and recording group interaction. The group interaction is directed by questions and topics supplied by the researcher. This automatically means that particular skills are required from the group interviewer (Fontana & Frey 1994).

The data from group interviews are the transcripts (or other records) of the group's interaction. They might be used as the only data-gathering technique in a study in conjunction with other quantitative or qualitative techniques.

Whatever may be the types of interviews, there are some practical aspects of interviewing that the interviewer has to keep in mind. Those practical aspects are as under:

(i) Interview respondents
(ii) Managing the interview
(iii) Recording

A brief discussion follows.

(i) *Interview respondents:* There are some issues in interviewing that are related to the respondents who are to be interviewed. The main issues are:

(a) Who are to be interviewed and why?

(b) How many respondents will be interviewed and how many times will each person be interviewed?

(c) When and for what period will each respondent be interviewed?

(d) At what place will each respondent be interviewed?

(e) How will access to the interview situation be organized?

If we pay attention to these five issues, it is obvious that the first two issues are related to the sampling plan for the research project, which itself depends to a greater extent, upon research questions and purposes. The next two issues are concerned with time and location that have impact upon quality of data. Recognition of the importance of these two issues enables the researcher or interviewer to take decisions that maximize the quality of data in light of ethical responsibilities towards those being interviewed. The last issue is concerned with gaining access. How it is done by the researcher depends on the particular research project and its setting and context. In fact, how the interviewers establish contact with the respondents and organize access affects the relationship between the interviewee and the interviewer at all stages.

(ii) *Managing the interview:* For managing interview successfully, the following five points must be kept in mind:

(a) Preparation of the interview schedule

(b) Establishing rapport at the beginning of the interview

(c) Communication and listening skills

(d) Sequence and types of questions to be asked

(e) Closing the interview

Experts are of the view that the importance of each of these points is determined by the nature and type of interview. For example, if the interview is a highly structured one, a schedule must be developed and pretested, because the quality of preparation will influence the quality of data. On the other hand, if the interview is an unstructured one, only a general awareness regarding the questions or topics is enough and would be kept deliberately open-ended and also, there would be no attempt to standardize it. In the case of unstructured interview, the communication skills in general and listening skills in particular are important one (Keats 1988; Minichiello et al. 1990). Minichiello et al. (1990) have reported 16 subskills of listening which they recommend a good interviewer must practice in order to improve their listening skills.

Asking questions is also important and has been analyzed extensively in research methods literature. The analysis includes the way questions are asked, the wording of the questions and the sequence and types of the questions that are asked. Patton (1990) has provided a classification of the questions to be asked according to the topics. According to him, questions may be classified into opinion/belief, experience/behaviour, knowledge, feeling, sensory and demographic background. He has also classified probe questions, which are often considered necessary at different stages during an interview. Finally, how the interview is closed is also important. Minichiello et al. (1990) have pointed out several important strategies for closing an interview.

(iii) *Recording:* While planning about the qualitative research to be done by the interviewer, it is essential to consider how interview data are going to be recorded. If the interview is a highly structured one, the recording of responses stands simplified because it is here only making checkmarks on a response sheet. However, for unstructured interview particularly open-ended interview, possibilities of tape recording, video recording and/or note taking clearly exist. Researches have shown that the situation dictates recording. If the interview is being conducted in the field, there may not exist good opportunity for electronic recording. Here the researcher will have to keep himself satisfied by note taking. Moreover, possibilities of using an appropriate method of recording have to be assessed in relation to practical constraints of the situation, cooperation of the respondent and type of the interview selected.

A review of the literature has revealed that researchers have used interview by adopting either of the two approaches or any one of them. One approach is called as *miner approach* and the other is called *traveller's approach* (Prakash 2011). In miner approach of interviewing, the researchers try to uncover the trust and facts lying with the respondents. By means of structured or semi-structured schedule, the interviewer tries to uncover the reality by digging deeper into the experiential world. In traveller's approach the interviewer constructs the major milestones covered during the process of the interview to delineate concepts, interpretations and inferences. Thus this approach comes nearer to what is known as *constructionist perspective*.

Indian psychologists, like western ones, have used interview to address problems of a wide variety in qualitative research. A few examples will illustrate this. Goyal (1997) conducted open-ended interview of middle-aged women in his study and reported that gynaecological difficulties were prominent and financial problems and daughter's marriage produced greater insecurity and loneliness among them. In a culturally changing village community (Khorwar tribe) Mishra (1997) used interview for investigating the role of family in health care and experience relating to health problems among women. His findings revealed that the major health problems comprised diarrhoea, child mortality, conjunctivitis, unwanted pregnancies and arthritis. Not only this, women were found to attribute physical, behavioural and supernatural causes for explaining their health problems. Vasuki and Reddy (1998) conducted open-ended interview of children in the age group of 9 to 15 years along with their parents. In this study, an attempt was made to describe parental perceptions of having a single child and children's perception of being a single child. Results revealed that majority of the parents as well as children appreciated the states of having/being a single child.

2. *Observation:* Observation has a long history in social sciences. For example, it has been extensively used by psychologists (Irwin 1980, Brandt 1981, Liebert 1995), by sociologists (Adler & Adler 1994) as well as by educational researchers (Foster 1996). When observation is used, it usually consists of *detailed notation* of events, behaviours and the contexts surrounding the events and observations. In observation practically all the senses such as *seeing, hearing, feeling* and *smelling* are integrated. Observation can be of physical environment, social interactions, nonverbal communications, planned and unplanned activities and interaction as well as of unobstrusive indicators. Observation is often referred to as *fieldwork* because it commonly takes place in a field.

In literature on observation as the technique of data collection, the terms *quantitative* an
qualitative are frequently used. Quantitative approach tends to be highly *structured* and require
some predeveloped observation schedules, usually in detailed format. Here the researcher has t
take a decision whether he will be using already existing observation schedules or he will b
developing an observation schedule for the purpose.

Qualitative approach to observation is much more *unstructured*. Here the researcher doe
not use any predetermined categories and classifications, rather he creates observations in
more natural and open-ended way. Irrespective of the technique of recording, the behavio
events are observed as they naturally unfold. The rationale here is that categories and concep
for describing and analyzing the observational data will tend to emerge later in the researc
during analysis rather than be brought to or imposed upon the data, from the beginning.

Patton (1990) has proposed *five* dimensions along which observations tend to vary. *First,* t
observer's role may vary from being a full participant to a complete outsider. When the observ
fully participates while doing observation, it is known as *participant observation,* which
commonly used in qualitative research. *Second,* the observer may do observation covertly, t

is, from behind a one-way mirror or with the knowledge of those being observed or with only some of those being observed being unaware of observation. *Third,* persons being observed may be given full explanations or partial explanation or false explanation regarding the purpose of observation. Sametimes no explanation can also be provided. The *fourth* dimension is duration. The observation may continue for a year or could be as brief as an hour. The *fifth* and final dimension is breadth of focus, that is, observation may vary from being quite broad to quite narrow.

The process of unstructured observation typically evolves through a series of different activities. It begins with selecting a setting and gaining access to it and then, starting the observation and reading it. As the study proceeds ahead, the nature of observation changes, typically sharpening in focus and leading to some clear-cut research questions that require selected observations. Silverman (1993) has proposed five different stages in organizing any unstructured observational study.

(a) Beginning the research where some general research questions are proposed
(b) Writing field notes
(c) Looking as well as listening
(d) Testing hypotheses
(e) Making broader links

Although observation may be nonparticipant or participant, in qualitative research, participant observation is more common. Participant observation can be understood as a process in two respects. *First,* the researcher must become gradually a participant and gain access to the individuals. *Second,* the observation must gradually move through a process that concentrates on the aspects of the research that are essential for providing solution to the research questions. Spradley (1980) has distinguished three phases of participant observation as under:

(a) *Descriptive observation:* This is the beginning phase of the observation where the researcher tries to develop an orientation to the field under study. This is used to grasp the complexity of the field and, at the same time, develop more concrete research questions and lines of vision.
(b) *Focused observation:* This is the second phase of the participant observation that narrows the perspective of the research on those processes and problems, which are considered relevant for the research questions.
(c) *Selective observation:* This phase hints towards the end of data collection and is focused on finding further evidence and examples for the types of practices commonly found in the second step.

There are some limitations of participant observations as used as a technique of data collection in qualitative researches. One limitation is that not all phenomena can be easily observed in situations. For example, biographical processes are difficult to be observed. In fact, events or practices that seldom occur (and considered crucial to the research question) can be captured only by luck or only by a careful selection of the situations of observation. Still another limitation is that the participant observation can hardly be standardized beyond a general research strategy and it does not make sense to see this as a goal for further methodological developments (Luders 2004).

Despite all these limitations, observation has been used in the Indian context for a variety of issues. For example, observational data have been used to understand the quality of health services. One study has demonstrated that the schedule caste (SC) and schedule tribe (ST) communities in villages experienced discrimination in health services provided at the household level (Babu et al. 2000). Likewise, Kar et al. (2000) conducted an observation study of dementia patients in which several behaviour problems were revealed. The observations of caregivers led the researchers to conclude that these patients were aggressive, verbally abusive and used to show socially inappropriate behaviour more often.

Overall, it can be said that observation, whether participant or nonparticipant, is often conceived as a part of other methods. Many researchers prefer to use observation as a technique of data collection along with other methods to complement the major research activities.

3. *Document review:* Documents, both historical and contemporary, are a rich source of data for social research. Documents may be understood as *standardized artifacts* available in particular formats such as notes, case reports, drafts, contracts, remark, diaries, annual reports, certificates, judgements, letters or expert opinions, biographies, autobiographies, institutional memoranda, reports and government pronouncements and proceedings (Jupp 1996; Wolff 2004). Researchers have classified this bewildering variety. One of the most popular classification of documents has been done by Scott (1990). He has distinguished 12 types of documents, which are constituted by a combination of two dimensions: *authorship* (who produced the document) and access to the documents. The authorship refers to the origin of the document (in three categories: *personal, official-private, official-state*) whereas access refers to the availability of documents to the individuals other than the authors (in four categories: *closed, restricted, open-archival* and *open-published*). This resulted in 3 × 4 = 12 categories of documents. For assessing the quality of documents, Scott has suggested four criteria which can be used for deciding whether or not a particular document or set of documents can be useful for the research.

(a) *Authenticity:* Is the document genuine and of unquestionable origin?
(b) *Credibility:* Is the document free from the error? Has it not been distorted?
(c) *Representativeness:* Is the document typical of its kind? If not, is the extent of its untypicality known?
(d) *Meaning:* Is the document clear and comprehensible?

There are some practicalities in using documents as the source of data collection. One useful suggestion is that the researcher should not start from a notion of factual reality in the documents and compare it to other sources of data collection. Since documents represent a specific version of realities constructed for specific purposes, they should be seen as a way of conceptualizing information rather than as 'information containers'. Another suggestion is that no part of the document should be taken as arbitrary and the researcher should start from the ethnomethodological assumptions of order at all points. This must also include the way a document is set up (Flick 2009).

In the Indian context, documents have been frequently used for conducting research in the field of leadership using documentary text. Sharma (1995) identified leadership issues as reflected in Indian thought depicted in the Ramayana, the Bhagavad-Gita and Dharmashastra in general. It was observed that the leader has to ensure that group members are committed to organizational objectives and has to coordinate the activities of group members. Finally, the leadership style is determined by the leader's own values, confidence in subordinates as well as management of uncertainty.

There are some limitations in using documents as source of data collection. Sometimes the necessary documents are not available, not accessible or stand lost. In some other cases, there may be people who can block access to the documents the researcher needs. One other problem associated with document analysis may be that sometimes the researcher has problems in understanding the content of the documents because he can't decipher the words, abbreviations, codes or reference that have been used because they are difficult to be read.

Besides these techniques of data collection, there are also some other techniques, which are used in qualitative research. According to Marhsall and Rossman (1999) data collection techniques like *questionnaire, surveys, projective techniques* and *psychological tests* can be used either in quantitative or qualitative data. Besides these, *proxemics* (study of people's use of space and its relationship to culture) and *kinesics* (study of body movements) are also used as qualitative data collection techniques. *Narratives* have also been used as qualitative data collection technique. *Narratives* are used to study the people's individual life stories. In fact, narratives are stories told by a sequence of words, actions or images and more generally the organization of information within those stories. Narrative, thus, allow researchers to approach the person's (interviewee's) experiential yet structured world in a comprehensive way. In narratives first the initial situation is outlined (that is, how things were started), subsequently, events relevant to the narratives are selected from the entire set of experiences and presented as a coherent progression of events (that is, how things developed) and finally the situation at the end of the development is presented (that is, what became) (Hermanns 1995).

In narratives, two methods are very common—*narrative interview* and *episodic interview*. In the narrative interview, the informant or interviewee is asked to present the history of an area of interest. The interviewer's task is to make the interviewee tell the story of the interest in question as a consistent story of all relevant events from beginning to end. In episodic interview, the question-answer sequences are combined with narratives (of episodes). Thus episodic interview yields context-related presentations in the form of narrative since these are very close to the experiences and their generative context than other presentational forms.

Thus we find that there are different methods of data collection in qualitative research. For better results, researchers generally use more than one method in the study. By selecting a complementary method, a researcher can cover the weaknesses of one method with strengths of another. Thus a good qualitative researcher often tries to include multiple methods of data collection. The use of multiple data collection techniques is technically known as *triangulation*, which permits verification and validation of qualitative data.

DATA ANALYSIS AND INTERPRETATION

The term 'data analysis' itself has different meanings among qualitative researchers. Data analysis, in general, means examining, sorting, categorizing, comparing, synthesizing, evaluating and contemplating the coded data as well as reviewing the raw and reviewed data. Since qualitative researchers mainly concentrate upon the study of social life in natural settings, and there are different ways of looking at and analysing social life, there are multiple perspectives and practices of analyzing qualitative data. Despite this variety, some experts have tried to identify the common features of qualitative data analysis. For example, Miles and Huberman (1994) have suggested that there are six moves common across different types of analysis. They have labelled their approach as *transcendental realism*. Tesch (1990) has pointed out that while no characteristics are common to all types of analysis, there are no fewer than twenty-six different approaches to the analysis of qualitative data. Likewise, Patton (2002) has also mentioned a number of sources for data analysis and interpretation. The use of computer software in qualitative data analysis is also becoming very popular.

Here an attempt would be made to discuss some specific techniques of qualitative data analysis. Some specific techniques of data analysis are as under:

1. Miles and Huberman's technique
2. Grounded theory analysis
3. Narrative and Hermeneutic analysis
4. Ethnomethodology and conversation analysis
5. Discourse analysis
6. Documentary and textual analysis
7. Use of computers in analysis of qualitative data

A brief discussion follows.

1. *Miles and Huberman's technique:* Qualitative data analysis by Miles and Huberman is a comprehensive system that is directed at tracing out lawful and stable relationships among social phenomena. Their analysis has three main components.

(a) Data reduction
(b) Data display
(c) Drawing and verifying conclusion

(a) *Data reduction:* Data reduction occurs throughout the analysis and it usually takes place in *three* stages. In the beginning stage, it occurs through editing, segmenting as well as summarizing the data. In the middle stage, it occurs through coding and memorizing as well as through its associated features such as finding themes, clusters and patterns. In the last and final stage, it proceeds through conceptualizing and explaining because developing any abstract concept is also one way of reducing data. In both qualitative as well as quantitative research, the objective of data reduction is to reduce the obtained raw data without significant loss of information. In qualitative analysis, an additional precaution regarding not losing information is not to strip the data from their context.

(b) *Data display:* There are different ways of displaying data. Graphs, charts, networks, diagrams such as Venn diagrams, causal model, etc., are the common modes of data display. Miles and Huberman consider data display important and essential at all stages of analysis because these displays enable data to be organized and summarized as well as they also show what distance the analysis has covered so far. Thus any good qualitative analysis involves repeated and interactive displays of data.

(c) *Drawing and verifying conclusion:* The major purpose of qualitative research is to draw conclusion from the data. In fact, the reasons for reducing and displaying data are to assist in drawing conclusion. Although drawing conclusion logically follows data reduction and data display, it in fact takes place at each stage of analysis. The researcher notes down the possible conclusions early in the analysis, but since they are vague and ill-formed, they are held tentative pending further work and sometimes sharpened during it. They are not formalized until all the data have been analysed, conclusions are obtained in form of propositions and once they are formed, they also need to be verified.

These three components are interwoven and interactive throughout the analysis. The above three components such as data reduction, data display and drawing conclusion present an overall view of data analysis and involve three main operations: coding, memorizing and developing propositions. Coding and memorizing are closely related and go together.

2. *Grounded theory analysis:* Previously we have defined grounded theory as both an overall approach to research as well as a set procedure for developing theory through analysis of data. Here grounded theory will be discussed only as a procedure for developing theory through analysis of data. The basic aim of grounded theory analysis is to generate abstract theory to explain what is contained in data. All of its procedures are oriented towards this aim.

The essential theme of grounded theory analysis is to find a core category at a high level of abstraction but grounded in data. Grounded theory analysis does this function in three steps. The first step is to locate conceptual categories in data and this is done at the first level of abstraction. At the second step, the researcher tries to find relationship among these categories. At the third stage, he conceptualizes and accounts for this relationship at higher level of abstraction. Obviously this means that there are three general types of codes: *substantive codes, theoretical codes* and *core codes*. Substantive codes are the initial conceptual categories in the data; theoretical codes connect these categories; and core codes are the higher order conceptualization of the theoretical coding around which theory is to be built up. Accordingly, there are three objectives of grounded theory analysis.

(a) The first objective is to find substantive codes in the data. Substantive codes are categories generated from the empirical data but at a more abstract level. In the first level of analysis, some of these substantive codes appear as more central in the data than others.

(b) The second objective is to find out relationship among the substantive codes using theoretical codes.

(c) The third objective is to find a high order, more abstract construct (that is, core category), which integrates these hypotheses into theory and which describes and explains them.

The process by which the first objective is achieved is called *open coding*, the second objective is achieved is called *axial coding* and third objective is achieved is called *selective coding*. These may not be done sequentially and therefore, they may be overlapping or done concurrently. A discussion of these codings is put forward.

Open coding is done at the first level of conceptual analysis. Here analysis begins by *fracturing* or *breaking open* the data. Here the researcher examines the data to condense them into some preliminary analytic categories or codes. The open coding necessarily involves a close examination of data (or some of the data), identifying the concept or conceptual categories inherent in data and theoretical possibilities that are carried by data. As we know, the concepts are the basic building blocks of theory. The purpose of open coding is to use the data to generate abstract conceptual categories for later use in theory building. In simple words, open coding means using the data to generate conceptual labels and categories for use in theory building. Researchers have shown that open coding is best done in a small group. This helps keep the grounded theory analysis on track as well as ensures theoretical sensitivity.

The process of labelling is guided by two main activities—*making comparison* and *asking questions*. By making comparison is meant that the different pieces of data are constantly compared with each other for generating abstract categories. Similar events and incidents are labelled and grouped to form categories. Asking question means one type of question in three forms is constantly asked. First, what is illustrated by this piece of data? or second, what does this piece of data represent? or third, what category or property of category is indicated by this piece of data? These two activities, that is, making comparisons and asking questions, focus directly on abstracting and raising conceptual level in the data.

In grounded theory labelling, in open code is also guided by what is called *concept-indicator model* (Glaser 1978). According to this model, a concept can have many different possible empirical indicators. When the researcher infers a concept from an indicator in the data, he, in fact, is abstracting—going upwards from a piece of empirical data to a more abstract concept. Since a concept has many indicators, the indicators are interchangeable with each other for the purposes of inferring the concept. It means that indicator 1 is the indicator of concept and indicator 2, indicator 3, etc., are also the indicators of the concept. The researcher compares one indicator with the other indicator, assesses their similarities and differences for inferring the concept. The researcher also asks constantly what is indicated by this piece of empirical data.

Axial (or *theoretical*) *coding* is the second stage in the grounded theory analysis. Here the main categories that have emerged from open coding of the data are interconnected with each other. Thus, it is a complex process of inductive and deductive thinking involving several steps because the researcher here organizes the codes, links them and discovers key analytic categories. The word 'axial' has been used by Strauss and Corbin and it indicates the idea of putting an 'axis' through the data. This axis, as they assumed, connects categories identified in open coding. However, Glaser (1978) has used a more general term 'theoretical coding' for describing this stage.

How is the work of interrelation carried out? To explain this interrelating work, some concepts that tend to connect things are needed. The connecting concepts are called *theoretical concepts*. That is the reason why Glaser has called it as theoretical coding. Strauss and Corbin (1997) have suggested a *coding paradigm model* to describe the set of concepts used for making connections between things. This model is based upon two axes: one goes from causes to phenomenon and to consequences and other one links context and intervening conditions as well as interactional strategies of the participants to the phenomenon (see Figure 22.1).

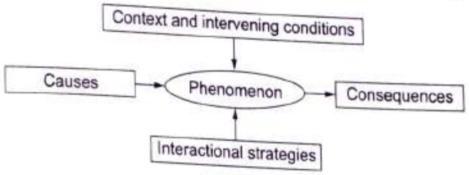

Figure 22.1 Coding paradigm model

The concepts included in each category may become (a) phenomenon for this category and/or (b) the context or conditions for other categories or a third group of categories, (c) a consequence. Thus the model names the possible relations between phenomena and the concepts and is used to facilitate the structure of relations between categories, between phenomena and between concepts. The developed relations and essential categories are repeatedly verified against the data and the related text. The researcher has to move continuously back and forth between inductive thinking (developing categories, concepts relations from the text, etc.) and deductive thinking (testing the concepts, categories and relations against text), especially against those passages or cases that are different.

From the different categories that were originated, only those are selected that seem to be most important and promising for further elaboration. These axial categories are considered fit with as many passages or cases as possible.

Selective coding is the third stage in grounded theory analysis. This stage continues the axial coding at a higher level of abstraction. The word 'selective' is used here because the researcher, for this stage, deliberately selects one aspect as core category or concept and concentrates upon it. At this stage, analysis begins around the core category and the core category becomes the central piece of the grounded theory. In a nutshell, in selective coding the researcher examines previous codes to identify and select data that tend to support the conceptual coding categories developed earlier. In selective coding, the primary objective is to integrate and pull together the developing analysis. The theory to be developed must have a central focus and this becomes the core category of the theory. It must be a central theme of the data and should also be perceived as central by the participants whose behaviour is being studied. In a nutshell, selective coding is thus aimed at developing the abstract, condensed and integrated picture of the data. For Strauss and Corbin, the important concepts here are core category and story line. The core category is defined as the central phenomenon around which other categories are integrated and story line means a descriptive narrative about this core category. When the storyline is analyzed, it becomes core category. In other words, in such analysis the researcher moves from a description of the storyline to a conceptualization of the story line. It means telling the story analytically, not descriptively and it also means finding a conceptual category to encompass what has been described in the story. Finally, theory is formulated in detail and again checked up against data. The procedure of interpreting data ends at the point where *theoretical saturation* has been arrived. Theoretical saturation means a stage where further coding, enrichment of categories, etc., no longer provide or promise any new knowledge. At the same time, the procedure is flexible in the sense that the researcher can re-enter the same text and same code from open coding with a different research question and can aim at formulating a grounded theory of a different issue. Thus this stage shows those categories where further data are required and therefore, directs further theoretical sampling. In the language of grounded theory, this stage is called as the *systematic densification and saturation of theory*.

3. *Narrative and Hermeneutic analysis:* Data analysis approaches based on coding and categorization are no doubt valuable in attempts to find and conceptualize regularities in data but the general limitation is that they break the data into small pieces and also generate the risk of the development of a culture of fragmentation (Atkinson 1992). In doing this, such approaches may decontextualize the data. Miles and Huberman have considered this problem as a serious one and have suggested ways of recombining and recontextualizing the data. Narratives and stories are such ways because they deal more holistically with qualitative data. Many qualitative data can be collected in story form as we find in case of oral and life histories, and biographical interviewing. Even where the data are not fully solicited in story form, they may come with stories characteristics where the participants may give narrative responses to the questions put forward by the interviewers. Narratives and stories have also been considered important in studying *lives* and *lived experience*. Narratives refer to the way people organize their everyday practices and subjective experiences, and they generally appear in oral or written texts for expressing the understandings. In fact, it is a quality of lived experience and a specific form by which people not only construct their identities but also locate themselves in whatever occurring around themselves, both at micro and macro levels. Any narrative shares six core elements: *telling a story or tale, sense of movement or process* (that is, before and after condition), *interrelating within a detailed context, an involved individual or collectivity that engages in action, coherence* (or the whole that holds together) and *temporal sequencing of the chain of events* (Neuman 2006). There are many research evidences in which stories have been used as a way to capture the lived experiences. For example, Brody (1987) and Coles (1989) have used stories for capturing lived experiences in case of medical illness; Riessman (1993) used this in studying major life events and trauma; and Delamont (1989, 1990) has used this in studies of education from students' point of view. This type of narratives, in fact, try to give a uniquely rich understanding of life situations and the story is often considered a feasible way of collecting data simply because it is considered as a very common device in social interaction.

Now the question is: how can qualitative data obtained from narratives and stories be analyzed? There are different suggestions from different experts. According to Schutze (1983), the first step in analyzing narrative interview is to eliminate all nonnarrative passage from the text and then to segment only purified narrative text for further analysis. At the second step, attempt is made to make a structural description of the contents where different parts of the narratives are specified. At the third step, the researcher moves away from the specific details of the life segments. Finally, the case analyses obtained in this way are compared and contrasted to each other. Bude (1984) has pointed out a different view by suggesting the reconstruction of life constructions. His viewpoint is that narratives take into account subjective and social constructions in what is presented, that is, life constructions in narrative interview, for example. Rosenthal and Fischer-Rosenthal (2004) have analyzed narrative interviews in five different steps.

(a) Analysis of biographical data (that is, data of events)
(b) Thematic field analysis (that is, reconstruction of the life story)
(c) Reconstruction of the life history (live as lived)
(d) Microanalysis of individual text segments
(e) Contrasting comparison of life history and life story

In narrative analysis, form and content can be studied together. Narratives can illuminate how informants use language to convey meanings and experiences. Here narrative analysis moves toward semiotics (Manning & Cullum-Swan 1994). It can be examined, for example, how language is used figuratively. Coffey and Atkinson (1996) have shown how analysis can explore participant's way of using imagery and how such devices as metaphors reveal shared meanings and understandings. The general exploration of using linguistic symbols for conveying shared cultural meanings is called as *domain analysis* (Coffey & Atkinson 1996). Metaphors used in social interaction for conveying meaning are also analyzed by qualitative analysts. Miles and Huberman have pointed out towards some useful properties of metaphors in qualitative analysis. For example, metaphors are used as data-reducing devices, decentering devices and ways of connecting findings to theory.

Hermeneutics is the study of interpretation of text or events in the humanities. Hermeneutical interpretation seeks to arrive at valid interpretations of the meaning of a text. This approach makes a distinction between (a) the subjective meaning that a statement or activity has for the participant and (b) its objective meaning that can be understood by using the concept of latent structure of meaning and this structure can be examined by using framework of multi-step scientific procedure of interpretation. That is the reason why it is also called as *objective hermeneutics*. Oevermann et al. (1979) have pointed out eight different levels of interpretation in objective hermeneutics.

(a) Paraphrasing the meaning of interactions according to verbatim text reported by the participants
(b) Explication of the interacting participant's intention
(c) Explication of objective motives of the interaction as well as of its objective consequence
(d) Explication of the function of the interaction for determining interactional roles
(e) Characterization of the linguistic features of the interaction
(f) Explication of the interpreted interaction for determining communicative figures
(g) Explication of general relations
(h) Independent test of the several hypotheses formulated at the preceding levels

There are three primary tools of narrative analysis: *path dependency, periodization* and *historical contingency*.

Path dependency refers to a process as having a beginning that triggers a structured sequence of events and create a path that is followed in the chain of subsequent events. The path limits the direction of the ongoing event that follows. In explanation based upon path dependency, the outcome is highly sensitive to events that occurred early in the process. Here the researcher starts with an outcome. He tries to demonstrate how the outcome follows from a sequence of prior events. As he traces back and displays the effect of each event upon the other, the researcher goes backward in the process to the initial event. Researchers using path dependency explanation assume that there may be one explanation for the starting event and another for the path of subsequent events.

There are two types of path dependency: *self-reinforcing* and *reactive sequence*. In self-reinforcing path dependency explanation the researcher looks at how an event once started, continues to operate at their own or tends to propel later events in a direction that resists external factors. Thus the initial trigger event places limits or constrains the direction of the process. The reactive sequence path dependency emphasizes upon a different process. Here the researcher focuses upon how each event responds to an immediately preceding one rather than upon tracing a process back to its origin. Thus here, the researcher very carefully examines each step in a process to see how one step influences the next one. Here the path does not necessarily have to be linear or unidirectional, rather it can bend or even sometimes reverse course to negate its previous direction. Thus here, the sequence of events can be like a pendulum, which swings back and forth. Any single event may set into motion a reaction that changes or may also reverse the direction of the events that preceded it.

Periodization is another important tool of narrative only where the researcher divides the flow of time in social reality into segment or periods. Thus here the researcher may for example, divide 150 years of history into periods by breaking continuous time in discrete units or periods, and define the periods theoretically. Theory helps the researcher in identifying what is significant and what is common within periods or between different periods.

In historical contingency, the researcher makes a unique combination of particular factors or specific circumstances that came together in a particular time and place. A contingent situation may be unexpected but once it occurs, it can profoundly influence subsequent events. Since there may be many possible idiosyncratic combinations of events, a researcher uses theory for identifying important contingent events for an explanation. Most of the time, researchers combine historical contingency and path dependency for the better results of the analysis.

4. *Ethnomethodology and conversation analysis:* Ethnomethodology is defined as a theoretical approach that is interested in analyzing the methods the individuals use in their everyday life to make communication and routine work effective. It addresses the question of how individuals produce social reality in and through interactive processes (Garfinkel 1967). In fact, ethnomethodology sets out to understand 'folk' (ethno) methods (methodology) for organizing the social world (Silverman 1993). The basic assumption of ethnomethodology is that people within a culture develop procedures for making sense of their daily life. The primary focus, therefore, is how the central features of a culture, its shared meanings and social norms are developed, maintained and changed rather than upon the content of those meaning (Feldman 1995). This focus forces ethnomethodologists to study activities that the ordinary people engage in often without proper thinking. For better understanding of joint action and interaction, language becomes central to the day-to-day activities. Thus social life is mediated through written and spoken communication where the study of language lies at the heart of ethnomethodology. In this way, conversation analysis becomes of central importance as ethnomethodologists seek to understand and interpret individual's methods for producing orderly social interaction. Heath and Luff (1996) have referred to a 1990 bibliography of ethnomethodological/ conversation-analytic studies that contain more than 1400 citations in five different languages. The general purpose of such studies is to understand the naturally occurring human conduct in which talk is the primary vehicle for production of human actions. Where only talk is analysed, verbatim transcripts of actual conversations are used.

Following Silverman (1993) the three fundamental assumptions of conversation analysis are: *structural organization of talk, sequential organization of tale* and *the need for empirical grounding of the analysis.* Based upon these assumptions, conversation analysis studies the situated production and organization of talk, developing an understanding of how context influences the participants' production of social reality.

Ten Have (1999) has outlined the following four steps involved in conversation analysis:

(a) Making recordings of natural interaction

(b) Transcribing the tapes either in whole or in part

(c) Analyzing the selected episodes

(d) Reporting the research

A frequent starting point for conversation analysis is to inquire into how certain conversations are opened and what linguistic practices are applied for ending these conversations in ordered way. There are two essential features of conversation analytic interpretation. First, conversation analytic interpretation is strictly a sequential procedure, that is, it ensures that no later statements or interactions are consulted for explaining a certain sequence. The turn-by-turn production of order in conversation is classified by conversation analysis. Second, conversation analytic interpretation emphasizes upon the context. This means that efforts in producing meaning or order in conversation can be analyzed only as related to the concrete contexts in which they are embedded in the interaction and in which interaction again gets embedded.

5. *Discourse analysis:* Discourse analysis is an important development in qualitative research. It includes not only everyday conversations but also other sorts of data such as

interviews or media reports. In fact, the word discourse captures this broader focus and refers to the general perspective within which ideas are formulated (Sapsford & Abbott 1996). Discourse analysis is very much sensitive to how spoken and written languages are used and how accounts and descriptions are constituted. At a microlevel, its elements share much with conversation analysis and therefore, some experts see conversation analysis as a particular type of discourse analysis (McCarthy 1991). At a macrolevel, the discourse analysis emphasizes upon the interrelationships between accounts and hierarchies, power and ideology. Discourse analysis is not an unified body of theory or methods, rather it is conducted within various disciplines with no overarching unifying theory common to all types. (Gee et al. 1992).

Despite this diversity and several disciplinary perspectives, there are some fundamental principles of discourse analysis.

(a) Human discourse is norm-governed and internally structured.

(b) It is produced by speakers whose cultural, economic, social and personal realities shape the discourse.

(c) Discourse itself contains important aspects of that socio-historical matrix.

Thus discourse analysis is concerned with any part of human experience constituted by the discourse.

At a more general level, Jupp (1996) has identified three features of discourse analysis as under:

(a) Discourse is a social process which indicates that words, sentences, etc., and their meanings depend on who used them where, why and to whom. Consequently, their meanings can vary according to social and institutional settings.

(b) There can be different types of discourse, which may be in conflict with each other.

(c) Besides being in conflict, discourse may be viewed as being arranged in hierarchy.

At a more specific level, Potter and Wetherell (1994) have also identified three features of discourse analysis that makes it relevant to the qualitative research.

(a) Discourse analysis is mainly concerned with talk and texts as social practices and therefore, it pays more attention to what has been called as linguistic content such as topics and meanings as well as attending to the features of linguistic form such as grammar and cohesion.

(b) Discourse analysis has shown concern with *action, construction* and *variability.* Individuals perform actions of different kinds through their talk and writing and they accomplish the nature of these actions partly by constructing their discourse out of a varieties of styles, linguistic resources and rhetorical techniques and devices.

(c) Discourse analysis is concerned with rhetorical or argumentative organization of talk and texts.

Potter and Wetherell (1994) have pointed out that it is very difficult to describe explicit procedures that are used in discourse analysis but they have identified five important considerations that are important. These are: variations as level, reading the details, looking for rhetorical organization, looking for accountability and cross-referencing discourse studies. There has been some differentiations in discourse analysis in the last four years. For example, Parker (2004) has developed a model of critical discourse analysis based upon the background developed by Foucault (1980). That is why this analysis is also known as *Foucauldian Discourse Analysis* (Flick 2009). Parker has suggested the following seven steps in such an analysis:

(a) The researcher should transform the text to be analyzed into written form, if it is not already done.

(b) The researcher should include free association to varieties of meaning for assessing cultural networks and these should be carefully noted down.

(c) The researcher should systematically itemize the objects in the text or in the selected portion of the text.

(d) The researcher should maintain a distance from the text by treating the text itself as the primary object of the study.

(e) The text should systematically itemize the subjects that is, characters, role positions, etc., in the text.

(f) The researcher should reconstruct the presupposed rights and responsibilities of the subjects.

(g) The researcher should map a network of relationships into patterns, which can be located in the relations of ideology, power and institutions.

6. *Documentary and textual analysis:* Analysis of various documents and related text is known as documentary and textual analysis. Such analysis shares characteristics with conversation and discourse analysis. But it has its own distinctive themes.

One important theme focuses on social production of the document and text, starting with how these came into being. All documentary sources are the result of human activity and are always located within the constraints of particular social, historical or administrative settings (Finnegan 1996). The words used and their meanings depend on where they are used and by whom to whom. Documents and texts studied out of their social context are actually deprived of their real meaning. Obviously, it means that an understanding of social production and context of the document and texts tend to have a strong impact upon its interpretation.

A second important and related theme is the social organization of document. Silverman (1993) has raised various questions relating to the study of social organization of documents irrespective of some errors involved therein. Such important questions are: How are documents written? For what purposes? On what occasions? Who writes them? What is recorded? What is omitted? What do readers need to know for making sense out of them?, etc. These questions tend to provide hints towards how documents and texts have been socially organized. Silverman has applied textual analysis to files, statistical records as well as records of official proceedings and images.

A third theme is concerned with a direct analysis of text and documents for meaning. Such analysis may focus on the surface meaning (or literal meaning), deeper meaning as well as upon multilayered nature of meaning, Historians often show concern with surface meaning whereas sociologists and psychologists are more interested in ways of uncovering deeper meaning. Methods used range from interpretive understanding to some structural approaches.

A fourth theme is the application of different theoretical perspectives to documentary and text analysis. According to Silverman, there are many ways of thinking about textual and documentary analysis and there are many different theoretical perspectives that can be applied.

7. *Use of computers in the analysis of qualitative data:* Qualitative research is undergoing technological change. A range of software programmes are available, mostly focused on the area of qualitative data analysis. Therefore, these programmes are sometimes called as QDA (Qualitative Data Analysis) software or as CAQDAS (Computer-Aided Qualitative Data Analysis Software). The various types of programmes or softwares available for analyzing qualitative data may be summarized as under:

(a) Word processors which allow the researcher to write as well as to edit text, and to search for words or word sequences

(b) Text retrieval programmes that allow researchers to search and summarize certain word sequences

(c) Code-and-retrieve programmes for splitting the text into segments and for retrieving or listing all segments of the text

(d) Text-based managers for administering, searching, sorting and ordering various text passages

(e) Code-based theory building that supports theory building by supporting steps and operations not only at the level of text but also at the conceptual level

(f) Conceptual Network Builders that help a researcher build and test theory by presenting graphic displays or network

Some of the examples of commonly used programmes are ATLAS.ti, NUD*IST/NIVO and MAXqda. ATLAS.ti was developed by Muhr (1991, 1994) and is based upon the approach of grounded theory and coding according to Strauss (1987). Its latest version is Nvivo7 and is only available for PCs. MAXqda was developed by Kuckartz (1995). One can create and import texts in Rich Text Format (rtf) from anywhere on one's hard disk and from the Internet by dragging and dropping. Some objects like PowerPoint slides, Excel tables and Photos may be imported as embedded objects of an rtf file.

These three programmes are just examples of the wide developing range of programmes and versions. More information regarding other programmes can be found at www.soc.survey.ac.uk/caqdas.

COMPARISON OF METHODS OF QUALITATIVE AND QUANTITATIVE DATA ANALYSIS

We have seen that there are different methods and techniques of data analysis in qualitative research. If we try to search what is the commonality among these various techniques, we shall arrive at the following general steps of data analysis:

1. The first step in analyzing qualitative research, irrespective of the methods, is *organizing the data*. The qualitative researcher often gets voluminous data from observations, interviews and/or documents. The method of organizing these data tends to vary depending upon the research strategies and data collection techniques used. For example, interview data are generally organized according to individual respondents whereas observation data may be organized individually or by grouping similar types of occurrences together. In fact, which approach will be taken by the researcher will depend upon the purpose of the research, the number of the participants, settings, time under study or on similarities and differences among the persons.

2. The second general step in data analysis in qualitative research is that of *description*. Here the researcher discusses different relevant aspects of the study including the viewpoints of the participants, setting both the temporal and physical purpose of any activity of these participant and impact of the activities of the participants, etc.

3. A final step in data analysis of the qualitative research is that of interpretation. According to Patton (1990), the interpretation involves explaining the findings by answering *why* questions and putting patterns into an analytic framework. However, the interpretation of qualitative research data is much dependent on the researcher's background, skills, biases and knowledge.

A good qualitative researcher also needs to keep in consideration the issue of internal and external validity of the research. Here internal validity is concerned with the accuracy of information and the way it matches with reality. Some qualitative researchers try to enhance the internal validity of their research by the process of triangulation. Likewise, external validity means that the researcher must keep a vigil on the generalizability of the obtained results and if possible, replicate his study and its generalizability.

In the previous section, various methods of analysis of qualitative data have been discussed. We have also seen the various methods of analysis of quantitative data in different chapters of the book. Do the two types of methods of analysis have some similarities? What are the basic differences between them? Let us see first the *similarities* and then the *differences*.

Similarities

1. In both qualitative and quantitative method of data analysis, the researcher very carefully examines empirical information to arrive at a conclusion based on reasoning and simplifying the complexity in obtained data. In fact, both forms of data analysis anchor statements about the social world.
2. In both qualitative and quantitative data, the comparison is a central process. Both types of researchers compare the evidences they have gathered internally or with other related evidences. By locating similarities and differences, both types of researchers try to identify multiple processes, causes or mechanisms within the evidence.
3. Both types of methods of analysis involve a public method or process. Both types of investigators collect data, describe them and document how they are collected and examined. However, the degree to which method is standardized may bring some variations but all researchers reveal their research design in some way or the other. In qualitative research, the research designs are not always explicit but they are implicit in every such research (King et al. 1994)
4. In both qualitative and quantitative method of data analysis, the researchers avoid false conclusions and misleading inferences. They remain alert for possible fallacies. They sort various explanations and discussions and try to adhere to the more authentic and valid ones.

Dissimilarities

The following are the major dissimilarities :

1. Quantitative data analysis is highly standardized and is built upon applied mathematics. Quantitative researchers choose from a specialized and standardized set of data analysis techniques. On the other hand, qualitative data analysis is less standardized. Qualitative research is often inductive and researchers rarely know the details of data analysis when they begin their project.
2. Quantitative researchers manipulate numbers, which represent empirical facts for testing a hypothesis with variable constructs. On the other hand, qualitative researchers tend to create new concepts and theory by blending together several abstract concepts. Qualitative researchers do not test a hypothesis rather they try to illustrate evidence, which shows that a theory or interpretation is plausible.
3. Quantitative researchers begin data analysis only when they have completed data collection and reduced them to some numerical facts. Subsequently, they manipulate the numbers for examining pattern or relationship. On the other hand, qualitative researchers look for a pattern or relationship during the early stage while they are still collecting data. In fact, the results of early data and analysis guide subsequent data collection.
4. These methods differ in the degree of distance from the details of social life. Quantitative researchers assume that social life can be easily measured by numbers, which can be manipulated by statistics to reveal features of social life. On the other hand, the data analysis in qualitative research does not depend upon well-established body of knowledge from mathematics and statistics. Here data are relatively diffuse and context-based and can have, generally, more than one meaning.

Thus we find that data analysis in qualitative and quantitative researches bear some resemblances and differences that mark their distinct status.

COMBINING QUALITATIVE AND QUANTITATIVE APPROACHES

Qualitative and quantitative research should not be considered as a mutually exclusive dichotomy, rather than as a continuum. In reality, for answering all of the questions some research studies need to include both qualitative and quantitative methods in the same research. A common area of research that uses both types of research methodologies is survey research. A single survey may contain questions that provide for quantitative responses and may also ask questions that might include qualitative data. Example of quantitative questions are:

How would you describe the performance of your local MLA?	5	4	3	2	1
	very good	good	neutral	poor	very poor

Example of qualitative question in the same survey would be:

How would you assess the overall performance of your local MLA?

Before paying attention to combining qualitative and quantitative approaches, let us look at the *central characteristics* of the two approaches.

Quantitative approach tries to understand reality in terms of variables and relationship among them. It is based on measurement and therefore, prestructured data, conceptual frameworks and design. It has well-developed methods for data analysis. Such methods are based upon the principles of statistics and mathematics. It places emphasis upon variable-oriented analysis. Its methods, in general, are unidimensional and less variable. Samples in quantitative research are larger than in qualitative studies.

Qualitative approach deals more with cases rather than with variables. It aims at an in-depth and holistic understanding for doing justice to the complexity of social life. Such an approach is sensitive to context, to live experiences as well as to local groundedness and a good researcher always tries to get closer to what is being studied. In qualitative research, samples are usually small and its sampling is guided by theoretical rather than probabilistic considerations. Here the researcher rarely prestructures the research design and data. Also, methods of analysis are less standardized. They are also more multidimensional, more diverse and less replicable.

Both quantitative and qualitative researches have their strengths. The central characteristics of the quantitative research discussed above show that quantitative researches enable standardized and objective comparison. Not only this, measurements involved in quantitative research permits a description of the situation or behaviour under study in a systematic and comparable way. Obviously then for the researcher, it becomes very easy to sketch the contours of these situations. The quantitative researcher also brings objectivity into the research. Here the researcher can systematically answer the important questions relating to the research and this automatically opens the way to the development of some useful knowledge.

Likewise there are some strengths of qualitative research, too. Qualitative research is more flexible than quantitative research and due to this flexibility, such research is well-suited for studying naturally occurring real-life situations. Qualitative research is considered the best means of studying the lived experience of the people including people's meanings and purposes. In other words, it is considered most appropriate for knowing about local groundness of the things as well as what is called actor's definition of the situation. Since qualitative data have the sense of holism and richness, they are well-suited for dealing with the complexity of social phenomena.

From the above description about the central characteristics and strengths of quantitative and qualitative research, two things are clear. First, the researcher needs to be very much clear regarding what he is trying to find out. Second, the researcher can't find everything he wants to know by using any one approach and therefore, he has to enhance the scope and power of research by combining both the approaches.

Now the question is why should quantitative approach and qualitative approach be combined? What is the justification for combining these two approaches? At the most general level, it can be said that the reason for combining these two approaches is to capitalize on the strengths of the two approaches as well as to compensate for the weaknesses or limitations of the two approaches. At the specific level, the reasons for combining these two approaches should be considered in the practical situation in light of the practical circumstances and context of the research.

Another question regarding combining these two approaches is what is the meaning of 'combining' here. Should the methods of the two approaches be combined? or Should the data of these two approaches be combined? or Should the methods, data and findings all be combined? There are studies, which have combined methods, data and findings of the two approaches. Such studies are called as *multi-method studies* (Brewer & Hunter 1989). Such studies have also been covered by Fielding and Fielding (1986). Besides multi-method studies, sometimes the term mixed method is also used to describe different types of combination of quantitative research and qualitative research.

Experts have pointed out that in considering how the two approaches are combined, some questions are important:

(a) Are the two approaches to be given equal weight?
(b) Are the two approaches separate or interactive?
(c) What is the logical relation of the two approaches? Are they only sequenced? or Are they really integrated in multi-methods design?
(d) What are the criteria used for evaluating the research? Is there a domination of a traditional view of validation or are both forms of research evaluated by appropriate criteria?

In fact, answering these questions and considering their implications, permits the researcher for using some sensitive design of qualitative and quantitative research in a reflexive way.

Different experts have written about combining these two approaches (Miles & Huberman 1994; Brannen 1992; Brewer & Hunter 1989; Bryman 1992 and Creswell 1994). A more comprehensive treatment has, however, been given by Bryman (1992) who has provided eleven ways of integrating quantitative and qualitative research. Therefore, we shall exclusively concentrate here on Bryman's approach.

1. *Logic of triangulation:* The findings of qualitative research can be checked against the findings of quantitative research. The purpose is generally to enhance the validity of results.
2. *Qualitative research can facilitate quantitative research:* Qualitative research can provide several background information regarding context and subjects, act as a source of hypotheses as well as facilitate scale construction.
3. *Quantitative research can support qualitative research:* It means that quantitative research may provide help regarding choice of subjects for a qualitative investigation.
4. *Quantitative and qualitative research are combined for providing a more general picture of the issue under study:* Quantitative research may be undertaken to plug the gaps in qualitative research that may arise, for example, due to the fact that the researcher can't make himself available at more than one place at a time. Alternatively, it is just possible that not all issues can be addressed solely to a quantitative research or solely to a qualitative research.
5. *Combining the strengths of qualitative and quantitative research:* Structural features of social life are analyzed with quantitative research and processual aspects are analyzed with qualitative approach. These strengths can be brought together in a single study.

6. *Researchers' and subjects' perspectives:* Quantitative research is usually driven by the researchers' perspective whereas qualitative research takes the subjects' perspective as the point of departure. These two emphases may be brought together in a single study.
7. *Problem of generality:* The addition of some quantitative evidences may help the researcher in mitigating the fact that only limited generalization is possible from qualitative research.
8. *Qualitative research may support the interpretation of relationships between variables:* Qualitative research permits the researcher for establishing the relationship among variables but seems often weak in explaining the relationship among variables. Qualitative research can, however, be used for helping in explaining the factors underlying the established relationship.
9. *Relationship between micro and macro levels:* When both qualitative and quantitative research are used in the same study, the micro-macro gulf is likely to be bridged. Quantitative research can often tap large-scale structural features of social life whereas qualitative research tends to tap small-scale behavioural aspects. When the researcher wants to explore both the levels, integrating quantitative and qualitative research becomes a wise step.
10. *Stage in the research process:* Qualitative and quantitative research may be appropriate to different levels of longitudinal research.
11. *Hybrids:* This form of research uses qualitative research in quasi-experimental design (quantitative research design).

Miles and Huberman (1994) have discussed four general designs which link qualitative and quantitative research design. In the first design, there is integrated collection of both types of data. In the second design, a multiwave survey is done in parallel with continuous fieldwork. In the third design, an exploratory fieldwork leading to quantitative instruments is conducted and quantitative data collection and analysis is followed by further qualitative work. In the fourth design also there occurs alternation where a survey is followed by in-depth qualitative work, which then leads to an experiment to test some hypotheses. Creswell (1994) has pointed out three designs which govern linkage of qualitative and quantitative research. The first design is called as *two-phase design* where quantitative and qualitative phases are separated from each other. The second is *dominant/less dominant design* where the researcher conducts a study within a single dominant paradigm with one small component drawn from the alternative paradigm. The third is the *mixed-methodology design* where some aspects of the two approaches are mixed at all the stages of the research.

☑ Review Questions

1. Give the meaning and essential features of qualitative research.
2. Explain Maxwell's model of qualitative research.
3. Trace the history of qualitative research
4. Discuss the major themes relating to design strategies of qualitative research.
5. Citing examples, outline the major theoretical perspectives of qualitative research.
6. Citing relevant examples, point out the major research design strategies of qualitative research.
7. Critically examine the different sampling techniques of qualitative research.
8. Discuss the major data collection techniques of qualitative research.
9. How does a qualitative researcher analyze and interpret data obtained from the research?

10. Give an outline of the important strategies for combining quantitative and qualitative research.
11. Make a comparative study of method of analysis of qualitative research and quantitative research.
12. Write short notes on the following:
    (a) Discourse analysis
    (b) Conversation analysis
    (c) Case studies
    (d) Document review
    (e) Grounded theory analysis

◆

23

CARRYING OUT STATISTICAL ANALYSES

CHAPTER PREVIEW

- Sample and Population
- Estimation, Confidence Intervals and Confidence Limits
- Normal Curve
    - Area under the Normal Curve
    - Non-normal Distribution: Skewness and Kurtosis
- Measures of Relative Position: Standard Scores
    - Linear Standard Scores
    - Normalized Standard Scores
- Parametric and Nonparametric Statistical Tests
- Parametric Statistics
    - Student's t and Z test
    - F Ratio
    - Pearson r
- Nonparametric Statistics
    - Chi-square Test
    - Mann–Whitney U Test
    - Rank-difference Methods
    - Coefficient of Concordance
    - Median Test
    - Kruskal–Wallis H Test
    - Friedman Test
- Correlation and Regression
- Major Terms and Issues in Correlation and Regression
- Choosing Appropriate Statistical Tests

SAMPLE AND POPULATION

In carrying out a statistical analysis it is important to distinguish between the two commonly used terms, namely, *population* and *sample*. A population may be defined as the totality of a particular characteristic for any specified group of individuals or objects. Thus, we may speak of a population of arts graduates, a population of science graduates, a population of science teachers, a population of medical books in the library, a population of university employees, etc. A sample may be defined as any selected number from a population. Usually, this selection is done according to some rule or plan. By studying the sample, some inferences are made about the population. A measure based upon a sample is known as a *statistic* and a measure based upon the population and inferred from a statistic is known as a *parameter*. A statistic, then, may be defined as a mathematical measure (based upon a sample) which helps in gathering, organizing, analyzing and interpreting the obtained data. Based upon the method of selection, a sample may be of two general types, namely, a probability sample and a nonprobability sample. The former is based upon a method in which chance plays a role and the latter involves a method in which samples are selected according to some judgement and the probability of the individual being selected is not known. A detailed discussion of probability and nonprobability samples has already been made in Chapter 14.

It is a popular misconception among students that a sample is a carbon copy of or has the identical properties of a population. In fact, when several samples are taken from the same population, they may vary from each other. But the nature of their variations, as pointed by the Central Limit Theorem, is reasonably predictable. Hence, a researcher, despite these variations, can make generalizations about a population on the basis of studying a sample. The Central Limit Theorem describes the nature of sample means and helps the researcher in making a generalization about the population with some known probabilities of error. The theorem states that if a large number of equal-sized samples (usually greater than 30 in size) is randomly taken from a population, the following three predictions will be true:

1. The means of the sample will be distributed normally.
2. The means of sample means will be identical with the mean of the population.
3. The distribution of sample means will have its own standard deviation which is technically known as the standard error of mean.

It can, therefore, be said that statistics are the most important tools in the hand of researchers, which enable them to make inferences or generalizations about a population with known possibilities of errors on the basis of observation of the characteristics of a population. Such inferences are known as *statistical inferences*. In any statistical analysis if we are dealing with the entire population, we use *descriptive statistics*. On the other hand, if we deal with a sample, our analysis relates not only to the characteristics of the sample but it also provides information about the population. Such statistics are known as *inferential statistics* or *inductive statistics* or *sampling statistics*.

Students should also know about the meaning of hidden population. By hidden population is meant a population of people who engage in clandestine, socially disapproved or concealed activities and who are really very difficult to be identified. To locate hidden population is considered very important in studies of deviant and stigmatized behaviour. For example, AIDS researchers have to draw samples from hidden populations. Many researchers often combine probability and nonprobability sampling for special situations such as for studying hidden population.

NORMAL CURVE

The concept of normal distribution is of utmost importance in statistical theory and practice, and no one is expected to interpret a statistic successfully without some understanding of the concept of normal distribution.

In the 18th century gamblers were interested in finding out the chances of beating an opponent at different gambling games and for this, they approached mathematicians. De Moivre in 1773 was the first such mathematician who developed a mathematical equation of the normal curve. Later on, Gauss and Laplace in the early 19th century rediscovered the normal curve independent of De Moivre's work. Gauss was primarily interested in the problems of astronomy, which led to the consideration of a theory of errors of observation. Hence, in that century the normal curve came to be known as the 'normal law of error'. In the middle of the 19th century, Quetelet promoted the applicability of the normal curve by popularizing the view that most of the problems in the field of anthropology, sociology, meteorology and human affairs could be solved with the help of the normal curve. In the latter half of the 19th century, Galton systematically undertook the study of individual differences and during this systematic study he found that most of the physical and psychological traits of human beings conformed reasonably to the normal curve. In this way, he extended the applicability of the normal curve. Today the normal curve is known by various names such as the *Gaussian Curve*, *De Moivre's Curve*, the *bell-shaped curve* or the *curve of error*.

A normal curve is one which graphically represents normal distribution. By definition, a normal distribution is one in which the majority of cases are located in the middle of the scale and only a small number of cases are located at both extremes of the scale. Most of the traits or variables in psychology, sociology and education are normally distributed. In reality, however, no distribution ever takes the absolute form of normal distribution. Many frequency distributions are close to this absolute form and we assume that they are normally distributed.

The major characteristics of a normal curve are enlisted below.

1. A normal curve is always symmetrical, that is, the right half of the curve is equivalent to the left half of the curve.
2. A normal curve is unimodal, and the mode is always at the centre of the distribution. In fact, in a normal curve the mean, the median and the mode are numerically identical and fall at the centre of the distribution.
3. A normal curve is asymptotic to the *x*-axis. Hence, a normal curve never touches the baseline, no matter how far the curve is stretched.
4. In a normal curve, the highest ordinate is at the centre. All ordinates on both sides of the distribution are smaller than the highest ordinate.
5. A normal curve is continuous.

Area under the Normal Curve

As we know, there is a definite relationship between standard deviation units and the normal curve. Figure 23.1 shows the different percentages of areas falling under a normal curve at different standard-deviation units. One standard-deviation unit taken on each side of the mean includes a total area of 34.13% + 34.13% = 68.26% of the curve. This is approximately two third of the cases. In terms of probability it can be said that in any normally distributed sample, chances are that two out of three scores will fall within the area of one standard-deviation unit on each side of the mean. A second standard-deviation unit taken beyond the first standard deviation cuts off 13.59% of the area on each side of the mean (Figure 23.1). Thus, up to two standard-deviation units on each side of the mean we have 68.26% + 27.18% = 95.44% of the total area. If we take another, or a third standard deviation, we have 2.15% of area on each side of the mean. The total areas included by all the six standard-deviation units (three standard-deviation units on each side), have accounted for 99.74% area of the normal curve. This means that only 0.26% of the cases are left which lie beyond three standard-deviation units from the mean. For convenience in statistical computation, we generally take and analyze cases up to three standard-deviation units on each side of the mean in normal distribution.

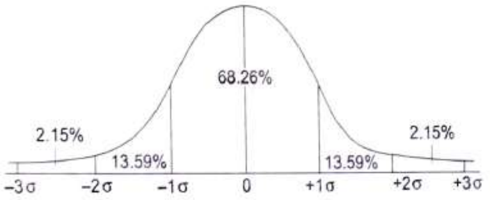

Fig. 23.1 Different percentages of area under the normal curve at various standard-deviation units

In psychological and educational researches, the normal curve has the main practical applications given below.

1. A normal curve helps in transforming the raw scores into standard scores.
2. A normal curve helps in calculating the percentile rank of the given scores.
3. If we want to normalize the obtained distribution, a normal curve is of immense importance.
4. A normal curve helps in testing the significance of the obtained measures against a chance hypothesis and thus enables the researcher to make a generalization about the population from which the sample was drawn.
5. The data obtained on the basis of the responses to attitude scales, ratings or rankings may be scaled in terms of the qualitative data by making suitable transformation in the numerical values.

Researchers in the field of psychology and education prefer a normal distribution of test scores, even though many other distributions are theoretically possible. Why? There are several reasons for it. Some of the important reasons are as under:

(a) One popular reason why researchers prefer normal distributions is that normal curve has some useful mathematical features that form the basis for several kinds of statistical investigation. Let us take an example to illustrate this fact. Suppose the researcher wants to know whether the average (mean) intelligence of first year medical students and first year engineering students of an institution were significantly different. An inferential statistic such as the *t* test for the difference between means would be the most appropriate statistic. However, this inferential statistic (many more also) is based upon the assumption that the underlying population of scores is normally distributed or nearly so. Thus for facilitating the safe and smooth use of inferential statistic, researchers prefer that the test scores in the population follow a normal or near-normal distribution.

(b) Another reason for preferring the normal distribution is its mathematical precision. Since the normal distribution is accurately defined in mathematical terms, it is possible to compute area under different regions of the curve with accuracy. For example, from normal curve we can easily determine that vast bulk of scores (more than 68 per cent) fall within one standard deviation of mean in positive and negative direction.

(c) The third reason for preferring a normal distribution of test scores is that normal distribution of test scores often arises spontaneously in nature. Due to this reason, early investigators considered normal curve as law of nature. In fact, important human characteristics, both physical and mental, tend to produce a close approximation to the normal curve especially when measurements for large and heterogeneous samples are graphed. Physical characteristics like birth weight, brain weight and mental characteristics like intelligence do have a near-normal distribution.

Nonnormal distributions: Skewness and Kurtosis

Nonnormal distributions are those distributions of score, which deviate from normal one. Nonnormality of distribution arises when some of the hypothetical factors determining the strength of a trait are dominant or prepotent over the others and therefore, are present more often than chance will allow. Two important types of nonnormal distributions frequently studied by psychologists and educational researchers are: *Skewness* and *Kurtosis*.

Skewness refers to the extent to which a frequency distribution departs from a symmetrical shape. If the test scores are piled up at the low (or left) end of the scale and are spread out gradually toward high or right end of the distribution, it is said to be *positively skewed*. On the other hand, when the test scores are plied up at the high or right end of the scale and are spread out gradually toward low or left end of the distribution, it is said to be *negatively skewed* (cf. Figures 23.2, 23.3). When the skewness is positive, the mean lies to the *right* of the median and when the skewness is negative, the mean lies to the *left* of the median.

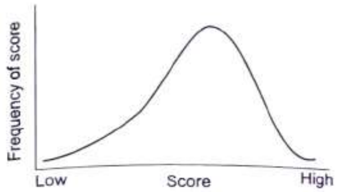

Fig. 23.2 Negative skewness

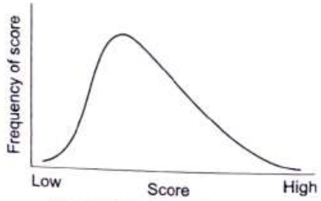

Fig. 23.3 Positive skewness

In normal distribution the mean equals the median exactly and therefore, skewness is zero. But in nonnormal distributions these two differ and accordingly, skewness is enhanced.

There are several formulas for calculating skewness. In terms of percentiles and median, skewness can be calculated as under:

$$S_k = \frac{\frac{P_{90} + P_{10}}{2} - Mdn}{D} \tag{23.1}$$

where, S_k = skewness; Mdn = Median; D = difference between the values of P_{90} and P_{10}.

Suppose for a distribution of intelligence scores, we get the following statistics:

$$Mdn = 80, \quad P_{10} = 20, \quad P_{90} = 150, \quad N = 100$$

then, $$S_k = \frac{\frac{150+20}{2} - 80}{150 - 20} = \frac{85 - 80}{130} = \frac{5}{130} = 0.038$$

In order to answer whether this distribution is skewed significantly or whether it may well be merely chance fluctuation from skewness of zero, we need to calculate its standard error. In this situation, the standard error of skewness is calculated by formula 23.2

$$S_{sk} = \frac{0.5185}{\sqrt{N}} \tag{23.2}$$

where, S_{sk} = Standard error of skewness.

By substituting, the value of skewness in formula 23.2, we get

$$S_{sk} = \frac{0.5185}{\sqrt{100}} = \frac{0.5185}{10} = 0.051$$

Dividing the skewness, that is, 0.038 by its standard error, that is, 0.051, $\left(\frac{S_k}{S_{sk}}\right)$, we get 0.745 as the value of *z*. Since this is smaller than 1.96, it is obviously not significant. Therefore, the obtained skewness 0.038 is not significantly different from zero. So, we conclude that there is *no* evidence to indicate that intelligence scores are not distributed symmetrically. This conclusion should be noted carefully and the erroneous inference that we have thereby proved that the scores are distributed symmetrically should be avoided. The fact to be noted is that the *skewness is not significantly different from zero does not justify in concluding that skewness is zero.*

Skewness can also be calculated in terms of mean, median and standard deviation as under:

$$S_k = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}} \qquad (23.3)$$

or,

$$S_k = \frac{\text{mean} - \text{mode}}{\text{standard deviation}} \qquad (23.4)$$

Its standard error, here, is given by

$$S_{sk} = \sqrt{\frac{3}{2N}} \qquad (23.5)$$

Now the basic questions is: *What do we do if a frequency distribution is skewed?* There are several answers to this question.

(i) The researcher computes skewness and its standard error. If the skewness is *not* significantly different from zero, he simply ignores it.

(ii) After computing skewness and its standard error, if the researcher finds that it is significantly different from zero, he thinks about it, looks at the distribution again and decides that the distribution is not very skewed but says or does nothing beyond reporting skewness and its standard error. This looks like a very casual procedure.

(iii) The researcher tends to alter the shape of the distribution, forcing it to become normal and thus, making $S_k = 0.00$. This process is called as normalization and has been widely used in determining *T* scores.

(iv) The researcher may transform the skewed distribution of scores by using some function of scores which will produce a distribution with little or no skewness. For example, if the distribution of *X* scores is positively skewed, such functions as $X_2 = \sqrt{x}$ (square root transformation such as $X = 2$ becomes $\sqrt{2} = 1.41$; $X = 6$ becomes $\sqrt{6} = 2.45$, and so on) will increase the spacing among the low scores and may reduce spacing among high scores. Thereby the value of skewness will be considerably reduced.

What are those situations that lead to skewness? Researchers have identified *five* such situations that generally produce skewness.

(i) There may be a natural restriction at the low end of scale. This leads to positive skewness. The examples are number of offspring, commission earned by life insurance agents, minimum speed of the car at stop signals, etc.

(ii) There may also be a natural restriction at high end of the scale resulting in negative skewness. Example is the number of rats that can be kept in a box.

(iii) There may also be artificial restriction at the low end of the scale and this may produce positive skewness. A good example is the distribution of scores earned by the examinees on an extremely difficult test.

(iv) There may also be artificial restriction at the high end of the scale resulting in negative skewness. Example is the distribution of scores earned by the examinees on an extremely easy test.

(v) Skewness may be found as a function of the measuring system. One example is a record of time required to complete a task when the researcher is really interested in speed.

How is skewness measure interpreted? When computed by formula 23.1, the maximum value of skewness can be +0.05 and the minimum value of skewness can be –0.50. Symmetry or lack of skewness is denoted by 0.00. If skewness is positive and $\frac{Sk}{S_{sk}}$ or *z* is + 1.96 or greater, we have obvious reason to conclude that the lopsidedness in our sample of cases is not just accidental and that the trait we are measuring is really positively skewed. It means that if we were to select some other group similar to the one used in the research, we would again find that deviations from mean or median will go in the positive direction rather than in the negative direction. However, if skewness is negative and if $\frac{S_k}{S_{sk}}$ is negative but somehow equal to or greater than –1.96, we may likewise conclude that trait is really negatively skewed. On the other hand whenever $\frac{S_k}{S_{sk}}$ lies between –1.95 and +1.95, it means that we have failed to prove that the skewness exists. Here two things are, however, to be clearly kept in mind: (a) this holds regardless of the value of skewness and (b) this never means we have established that the curve is symmetrical.

In psychological and educational testing, skewed distributions usually signify that the test constructor has included too few easy items or too few hard items. For example, when scores are concentrated at the low end (positive skewness), it means that the test contains too few easy items for making effective discriminations at the end of the scale. In such a case, the examinees who have obtained zero or near-zero score might actually differ with respect to the dimension measured. But the test is not able to find out these differences because most of the items are too hard for the examinees. On the other hand, if scores are massed at the high end (negative skewness), it means that the test contains too few hard items for making effective discrimination at this end of the scale.

The most direct solution to such skewness is to add items or modify the existing items so that the test has more easy items (for reducing positive skewness) or more hard items (for reducing negative skewness). If it is too late to revise the test items, the test constructor can use a statistical transformation to help produce a more normal distribution of scores. However, the most preferred good strategy is to revise the test so that the skewness is kept at the minimum or it becomes nonexistent.

Kurtosis refers to the peakedness or flatness of a frequency distribution as compared with normal distribution. In simple words, kurtosis is the extent to which a curve is more or less peaked than a normal curve with the same standard deviation. When a frequency is more peaked than the normal one, it is said to be *leptokurtic*; when flatter than normal, it is called as *platykurtic* and when the frequency distribution is neither peaked nor flat-topped but as the same as the normal curve, this medium kurtosis is described as *mesokurtic* (see Figure 23.4).

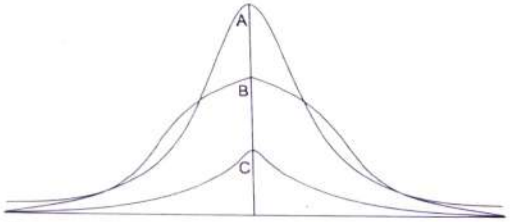

Fig.: 23.4 Types of kurtosis: curve A is leptokurtic; curve B is mesokurtic; curve C is platykurtic

In Figure 23.4 curve *A* is leptokurtic, curve *B* is mesokurtic and curve *C* is platykurtic. Measures of kurtosis are not widely used. Hence a simple formula for calculating kurtosis is in terms of percentiles given on the next page.

$$K_u = \frac{Q}{P_{90} - P_{10}} \qquad (23.6)$$

where, K_u = Kurtosis;

Q = Quartile deviation

The numerical value of *Ku* must always be between 0.00 and 0.50 although most values will fall within the limits of 0.21 and 0.31. For normal distribution, formula 23.6 gives $Ku = 0.263$. If *Ku* is greater than 0.263, the distribution is platykurtic and if less than 0.263, the distribution is leptokurtic. Kurtosis can also be calculated by another formula (Downie & Heath 1970).

MEASURES OF RELATIVE POSITION: STANDARD SCORES

The relative position of a score means its distance from the mean, expressed in terms of some deviational measures such as standard deviation. A standard score, which is a kind of derived score, is a method of expressing the distance of the score from the mean in terms of standard deviation. The 'standard' about a standard score is that it has a *fixed mean* and a *fixed standard deviation*. Standard scores can be classified into two most common categories.

Linear Standard Scores

The underlying purpose of transforming any original scores into standard scores is to make the scores on different tests comparable. There may be linear or nonlinear transformation of the original scores. A linear standard score is one where linear transformation of original scores is made. When standard scores are based upon linear transformation, they retain all the characteristics of original raw scores because they are computed by subtracting a constant (such as mean) from each raw score and then dividing the obtained value by another constant (such as *SD*). Since all characteristics of original raw scores are duplicated in linear standard scores, any statistical computation that can be done with original raw scores can also be done with such standard scores. The most common examples of linear standard scores are the sigma scores (or *z* scores), Army General Classification Test (AGCT) scores, College Entrance Examination Board (CEEB) scores and Wechsler Intelligence Scale DIQs. These linearly derived standard scores can be compared among themselves only when they are obtained from distributions which have approximately similar shapes.

A sigma score is one which expresses how many standard-deviation units a particular score falls above or below the mean. To compute this, we subtract the mean from each original score and then, divide the result by the standard deviation. A detailed discussion of the sigma score has already been done in Chapter 7. As discussed in that chapter, a *z* score or sigma score has two important limitations, namely, occurrence of negative values and occurrence of decimal fractions, which make a sigma score difficult for use in further statistical calculation as well as in reporting. To get rid of these difficulties, further linear transformation of sigma scores is made. AGCT scores, CEEB scores and DIQs in Wechsler Intelligence Scales are examples of such linear transformation. The AGCT scores employ a mean of 100 and standard deviation 20; CEEB scores employ a mean of 500 and standard deviation of 100; and WISDIQs employ a mean of 100 and a standard deviation of 15 for further linearly transforming the sigma scores. To convert the original sigma score into any of the above standard scores, we simply need to multiply the standard score by the desired SD and add or subtract it from the desired mean value.

Normalized Standard Scores

Sometimes researchers may wish to compare the scores obtained from distributions having dissimilar shapes. In such situations, they employ some nonlinear transformations so that the scores may conveniently fit into any specified type of distribution, preferably the normal distribution. The reason for choosing a normal distribution is twofold. First, most of the characteristics encountered in behavioural researches are normally distributed among the population and second, a normal distribution facilitates further statistical computation. Thus, the normalized standard scores may be defined as those standard scores which have been expressed in the form of a distribution that has been transformed in a way that fits a normal curve. Like linear standard scores, normalized standard scores can be expressed in terms of the mean of zero and standard deviation of 1. If a person has obtained zero as a normalized standard score, it indicates that the performance of that person lies exactly at the mean and hence, he excels over 50% of the persons in his group. Likewise, if a person's performance is at +2SD units above the mean or normal curve, it means he surpasses about 98% of the persons in his group. Likewise, if his performance falls at –2SD units below the mean, it means he surpasses only about 2% of persons in his group. The common examples of normalized standard scores are the *T*-score and Stanine score, which have already been discussed in detail in Chapter 7. The reader should note carefully that if the distribution of original scores is a normal one, the linear standard scores and the normalized standard scores would yield more or less identical results. A normalized standard score is preferred only if the situation meets the following requirements:

1. The sample is large.
2. The sample is representative.
3. The non-normality of the distribution of original scores is not due to the behaviour or trait under consideration, rather it is due to some defects in the test material itself.

PARAMETRIC AND NONPARAMETRIC STATISTICAL TESTS

The parametric and nonparametric statistical tests are commonly employed in behavioural researches. A parametric statistical test is one which specifies certain conditions about the parameter of the population from which a sample is taken. Such statistical tests are considered to be more powerful than nonparametric statistical tests and should be used if their basic requirements or assumptions are met. These assumptions are based upon the nature of the population distribution as well as upon the type of measurement scales used in quantifying the data. The assumptions may be enumerated as follows:

1. The observations must be independent. In other words, the selection of one case must not be dependent upon the selection of any other case.
2. The observations must be drawn from a normally distributed population.
3. The samples drawn from a population must have equal variances and this condition is more important if the size of the sample is particularly small. When the different samples taken from the same population have equal or nearly equal variances, this condition is known as *homogeneity of variance*. Statistically speaking, by homogeneity of variance is meant that there should not be a significant difference among the variances of different samples.
4. The variables must be expressed in interval or ratio scales. Nominal measures (that is, frequency counts) and ordinal measures (that is, rankings) do not qualify for a parametric statistical test.
5. The variable under study should be continuous.

The examples of a parametric test are the *z* test, *t* test and *F* test.

A nonparametric statistical test is one which does not specify any conditions about the parameter of the population from which the sample is drawn. Since these statistical tests do not make any specified and precise assumption about the form of the distribution of the population, these are also known as *distribution-free statistics*. The nonparametric statistics do not specify any rigid conditions like parametric statistical tests although certain assumptions are associated with them. For a nonparametric statistical test, the variables under study should be continuous and the observations should be independent. But these assumptions are neither rigid nor so elaborate as we find in the case of a parametric statistical test. The examples of nonparametric tests are the chi-square test, the Mann–Whitney *U* test, Kendall's tau, Kendall's coefficient of concordance, etc. A nonparametric statistical test should be used only in the following cases:

1. The shape of the distribution of the population from which a sample is drawn is not known to be a normal one.
2. The variables have been quantified on the basis of nominal measures (or frequency counts).
3. The variables have been quantified on the basis of ordinal measures (or ranking).

Because nonparametric statistical tests are based upon frequency counts or rankings rather than on the measured values, they are less precise, and are less likely to reject a null hypothesis when it is false. That is why, a nonparametric statistical test is used only when the parametric assumptions cannot be met. Some statisticians, however, argue that nonparametric statistical tests are more powerful and have more merits than parametric tests because their validity is not based upon the assumptions about the population distribution. They further argue that the parametric assumptions are often ignored by the researchers and there are evidences in certain parametric statistical tests like the *t* test and *F* test that violation of assumptions, particularly when a sample is large, does not affect the power of the statistical tests (Edward 1968; Winer 1971). Not only this, for some population distributions nonparametric statistical tests are superior in power to parametric statistical tests (Whitney 1948).

Table 23.1 presents a summary of the levels of quantitative description and types of statistical analysis appropriate for each level of measurement.

Table 23.1: Levels of quantitative description of parametric and nonparametric data

Level	Scale	Process	Data	Important appropriate statistics
1.	Nominal scale	Classified and counted	Nonparametric	Mode, chi-square test and sign test
2.	Ordinal scale	Rank in order		Median, *Q*, Stanine, Spearman's rho, Kendall tau, Mann–Whitney *U* test, Wilcoxos Sign Rank test, Kendall's partial rank order correlation
3.	Interval scale	Equal Intervals, No true zero point	Parametric	Mean, Standard deviation, Pearson *r*, *t* test, ANOVA, ANCOVA, Factor analysis
4.	Ratio scale	Equal intervals, True zero, Ratio relationship	—	—

Students should keep in mind the fundamental differences between parametric and nonparametric tests.

- Parametric tests rest upon some specific assumptions about population parametric whereas nonparametric tests don't make such explicit assumptions.
- In parametric test, the measure of central tendency is *mean* whereas in nonparametric test the measure of central tendency is *median*.

- In parametric test, there is complete information about the population whereas in the nonparametric test, there is no information about the population.
- The parametric tests are applicable to variables only whereas the nonparametric tests are applicable both to variables and attributes.
- In parametric test, the measurement of variables is done on *interval or ratio* level whereas in nonparametric test the variables are measured on *nominal or ordinal* level.
- In parametric test, the test statistic is based upon the distribution whereas in nonparametric test, the test statistic is arbitrary.

However, it should be made clear that many parametric statistics such as *t* test, ANOVA, ANCOVA, Pearson's *r* are considered still appropriate even when the assumption of normality has been violated. This robustness has been demonstrated especially for *t* test, ANOVA, and ANCOVA by researchers like Mandeville (1972), and Glass, Packham and Sanders (1972). Not only this, these statistical procedures which are suited for interval data and ratio data, can be applied to ordinal data as well as to dichotomous data (such as pass-fail, male-female, etc.) (Best & Khan 2006).

Bradley (1968) has enumerated several advantages and disadvantages of nonparametric statistical tests in comparison to parametric statistical tests. Some of his important advantages are as follows:

1. *Simplicity and facilitation in derivation*: Most of the nonparametric statistics can be derived by using simple computational formulas. This advantage does not lie with most of the parametric statistics, the derivation of which requires an advanced knowledge of mathematics.

2. *Wider scope of application*: Since nonparametric statistics as compared to parametric statistics are based upon fewer and less rigid and elaborate assumptions regarding the form of population distribution, they can be easily applied to much wider situations.

3. *Speed of application*: When the sample size is small, calculation of nonparametric statistics is faster than parametric statistics.

4. *Susceptibility to violation of assumptions*: In case of nonparametric statistics, assumptions are fewer and less elaborate than in the case of parametric statistics. Therefore, assumptions of nonparametric statistics are less susceptible to violation. Not only this, these violations are easier to check and can be readily and economically taken care of with the nonparametric statistics.

5. *Type of measurement required*: Nonparametric statistics require measurement based upon a nominal scale and ordinal scale, whereas parametric statistics require measurement based upon the interval scale and/or ratio scale. As treatments associated with either nominal scale or ordinal scale are easier than treatments associated with either interval scale or ratio scale, the parametric statistics have a better case for applicability than the nonparametric statistics.

6. *Impact of sample size*: When sample size is 10 or less than 10, nonparametric statistics are easier, quicker and more efficient than the parametric statistics. If the assumptions of parametric statistics are violated for such small cases, the result is likely to get badly affected. Therefore, for this sample size, nonparametric statistics are always superior to the parametric statistics. The reader should note that as the sample size increases, nonparametric statistics become *time-consuming, labour-intensive* and *less efficient* than the parametric statistics.

7. *Statistical efficiency*: Nonparametric tests are often more convenient than the parametric tests. If the data is such that it meets all assumptions of nonparametric statistics but not of parametic statistics then nonparametric statistics have statistical efficiency equal to parametric statistics. If both parametric and nonparametric statistics are applied to the data which fulfils all assumptions of parametric tests, the distribution-free statistics become more efficient with a small sample size but they become less and less efficient as the sample size increases.

Despite the above advantages (which, in general, hint towards disadvantages of parametric statistics), certain disadvantages are associated with them. The main ones are as given below.

1. According to Moses (1952), the nonparametric statistics have lower statistical efficiency than parametric statistics when sample size is large, preferably above 30.
2. If all assumptions of parametric statistics are fulfilled, Siegel (1956) and Siegel & Castellan (1988) consider the use of nonparametric statistics as simply 'wasteful of data'.
3. It is also said that the probability tables for testing the significance of nonparametric statistics are widely scattered in different publications which, for a behavioural scientist, is difficult to locate and interpret.

Where both parametric statistics and nonparametric statistics are available, the following guidelines should be kept in view for selecting any one of the two:

(i) If possible and the situation permits, parametric statistics should be used because they take the amount of difference between cases into account rather than simply the order of the cases.
(ii) The central limit theorem clearly ensures that parametric statistics will prove beneficial even for non-normal distributions.
(iii) If the data are available from non-normal distribution, the same can be transformed so that parametric statistical tests can be used (Hollander et al. 2014)
(iv) For larger samples ($N > 30$), parametric tests can be safely used.
(v) If the sample size is very small, there is no alternative to using a nonparametric statistic.

Now, we shall discuss some important parametric and nonparametric statistics commonly used in behavioural researches.

PARAMETRIC STATISTICS

The most important parametric statistics are

1. Student's *t* test and *z* test
2. Analysis of variance: *F* ratio
3. Analysis of covariance
4. Pearson *r*
5. Partial correlation and Multiple correlation

A detailed discussion of each of them is presented below.

1. Student's *t* test and *z* test

When the researcher wants to test the significance of difference between two means, he uses either the *t* test (or *t* ratio) or *z* test (or *z* ratio). The computation of *t* or *z* involves the computation of a ratio between the experimental variances (that is, the obtained difference between two means) and the error variance (that is, standard error of the mean difference). However, there are two basic differences between t ratio and *z* ratio. When the sample size is less than 30, we use the *t* test or Student's *t* for testing the significance of the difference between two means. This concept of small sample size test was developed in 1915 by William Seely Gosset, a statistician for Guinness Breweries in Dublin, Ireland. Because the service code prohibited publication under a researcher's name, he signed the name 'student' for publication of this test. Hence, this statistic is also known as Student's *t*. When the sample size is more than 30, the ratio of the difference between two sample means to the standard error of this difference is calculated by the *z* ratio which is interpreted through the use of normal probability tables. Another difference is that *z* test uses the actual population standard deviation whereas *t* test uses the sample standard deviation as an estimate when actual population standard deviation is not known. The reader should note that as the sample size increases, the critical values of *t* necessary for rejecting the null hypothesis gradually reduce and approach the *z* values of the normal probability table. There are some basic assumptions of *t* test. *First*, the variances of the two samples are nondifferent, that is, there exists homogeneity in variances of the samples. *Second*, the sample has been randomly selected. This condition helps to ensure that the sample is representative of the population so that accurate generalizations can be done from sample to population. *Third*, the population distribution of scores is normal. In fact, this requirement stems from a unique property of normal distribution. When it is stated that the observations are independent and random, the sample means and sample variances can be independent only when the population is normally distributed. *t* ratio as the test of significance of means demand that the means and variances must be independent of each other. They can't vary together in some systematic way—that is, one (mean) becoming larger with another (variance) also becoming larger. If these tend to vary in a systematic way, the underlying population cannot be normal and test of significance based on the assumption of normality would be treated as invalid. It is for this reason the assumption of normality lies behind the use of *t* test.

Student's *t* test is referred to as a *robust test*, which means that statistical inferences are likely to be valid even when there are large departures from normality in population. The *t* test is likely to be robust to the violations of normality when large samples ($N > 30$) are used. Therefore, in case of any serious doubts concerning the normality to population distribution, it is wise to increase *N* in each sample (Elifson 1990). Before discussing the small sample *t* test, five important concepts, namely, *degree of freedom, null hypothesis, level of significance, one-tailed test vs two-tailed test and power of test* should be considered.

Degree of freedom

The degree of freedom means freedom to vary. It is abbreviated as *df*. In statistical language, it can be said that the degree of freedom is the number of observations that are independent of each other and that cannot be deduced from each other. Suppose we have five scores and the mean of five scores is 10. The fifth score immediately makes adjustment with the remaining four scores in a way which assures that the mean of all five scores must be 10. For example, suppose we have four scores 12, 18, 5, 12, and then the fifth score must be 3 so that the mean becomes 10. In another distribution if the four scores are 2, 10, 8, 5, the fifth score must be 25 in order to have a mean of 10. The meaning is that four scores in the distribution are independent or they may have any value and they cannot be deduced from each other. The size of the fifth score, however, is fixed because the mean in each case is 10. Hence $df = N - 1 = 5 - 1 = 4$. Take an example of larger cases. Suppose we have a set of 101 scores. We compute the mean and in computing the mean, we lose 1 *df*. We had initially 101 *df* (because there were 101 scores) but now after computing mean, we have $N - 1 = 101 - 1 = 100$ degrees of freedom. Sometimes we have paired data and in such cases, the number of degrees of freedom is equal to one less than the number of pairs.

Null hypothesis

The starting point in all statistical tests is the statement of null hypothesis (H_0), which is a *no difference hypothesis*. In other words, a null hypothesis states that there is no significant difference between the samples under study. It makes a judgement about whether the obtained differences between samples are due to some true differences or to some chance errors. The null hypothesis is formulated for the express purpose of being rejected because if it is rejected, the alternative hypothesis (H_1) which is an operational statement of the investigators' research hypothesis, is accepted. As we know, a research hypothesis is nothing but predictions or deductions drawn from a theory. The tests of the null hypothesis are generally called tests of significance, the outcome of which is stated in terms of porbability figures or levels of significance.

If the difference between the experimental group and the control group is very small, the experimenter is likely to accept the null hypothesis, indicating the fact that the small difference between these two groups is due to sampling errors or some other chance fluctuations. On the other hand, if the difference between the experimental group and the control group is too large, the experimenter is likely to refute or reject the null hypothesis, indicating the fact that the obtained differences are real differences between or among the samples under study.

Level of significance

As stated above, the null hypothesis is developed for the express purpose of rejection. The rejection or acceptance of the null hypothesis is based upon the level of significance, which is used as a criterion. The levels of significance are also known as *alpha levels*. In psychological, sociological and educational researches there are two levels of significance which are commonly used for testing the null hypotheses. One is the 0.05 or 5% level and another is the 0.01 or 1% level of significance. If the null hypothesis is rejected at the 0.05 level, it means that 5 times in 100 replications of the experiment, the null hypothesis is true and 95 times this hypothesis would be false. In other words, this suggests that a 95% probability exists that the obtained results are due to the experimental treatment rather than due to some chance factors. Rejection of the null hypothesis when, in fact, it is true, constitutes a *Type I error* or *alpha error*. Thus it can be said that at the 0.05 level of significance, the experimenter commits a 5% Type I error when he rejects a null hypothesis. Some investigators may want a more stringent test and the 0.01 level is one such level where the investigator commits a Type I error of 1% only. The 0.01 level suggests that a 99% probability exists that the obtained results are due to the experimental treatment, and hence, once in 100 replications of the experiment, the null hypothesis would be true. Sometimes the investigator wants even more stringent test of significance and for this, he chooses the 0.001 level, which is uncommonly used in behavioural researches. This level suggests that only once in 1000 replications of the experiment, the null hypothesis would be a true one and in 999 replications (out of 1000) the obtained results can be attributed to the experimental treatment. In testing the significance of the obtained statistics, sometimes the investigator accepts the null hypothesis when, in fact, it is false. This error is technically known as the *Type II error* or *beta error*.

It is obvious from the above interpretation that an error of Type I can be reduced by putting the alpha level at the 0.01 or 0.001 level. But as we reduce the chance for making a type I error, we increase the chance level for making a Type II error where we do not reject the null hypothesis when it should be rejected. Therefore, as we decrease the possibility of making one type of error, we also increase the probability of making another type of error. The research workers must be *cautious* in this situation and should try, as far as possible, to limit the probability for making a Type I error.

What is the relationship between Type I and Type II errors? Researchers are of the view that when it comes to setting significance levels, protecting against one kind of error leads to the chance of making the other. The common insurance policy against Type I error (that is, lowering the alpha level from 0.05 to 0.01 or 0.001) has the cost of increasing the probability of committing Type II error. This happens probably because with a stringent significance level like 0.001 even if the research hypothesis (H_1) is true, the statistical results must be quite strong to be extreme enough to reject null hypothesis (H_o). The safeguard or insurance policy against Type II error (setting alpha level of say 0.30) has the cost of increasing the chance for making Type I error. This is because with the level of significance such as 0.30, even, if the null hypothesis is true, it is easy to get a significant statistical result enough for rejecting it. Thus the trade-off between these two conflicting concerns usually is put to rest by a compromise—formulating the standard 0.05 (5%) or 0.01 (1%) significance levels.

The whole issue of possibly correct or mistaken conclusions in hypothesis testing has been presented in Table 23.2.

Table 23.2 Statistical decision making: Possible correct and incorrect decisions

	Null hypothesis (H_o) True	Null hypothesis (H_o) False
Fail to reject H_o	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

At the top of Table 23.2, there are two possibilities: H_o is true and H_o is false. Likewise along the left side of the table, there are also two possibilities: Fail to reject H_o and Reject H_o. Table 23.2 clearly shows that there are two ways to be correct and two ways to be in error in any hypothesis-testing or decision-making situation.

One-tailed test vs two-tailed test

One-tailed test is a directional test, which indicates the direction of the difference between the samples under study. Suppose the experimenter conducts an experiment in which he takes two groups—one is the control group and the other is the experimental group. Only the experimental group is given training for five days on various kinds of arithmetical operations. Subsequently, an arithmetical ability test is administered on the two groups and the scores are obtained. In the above situation, the experimenter has reason to say that the mean arithmetical score of the experimental group will be higher than the mean arithmetical score of the control group. This is the alternative hypothesis, which indicates the *direction of difference*. When the alternative hypothesis states the direction of difference, it constitutes a one-tailed test. The null hypothesis would be that the mean of the experimental group is equal (no difference) to the mean of the control group. If this is rejected, we accept the above alternative hypothesis.

Putting the above facts schematically, we can say

1. $H_o = M_1 = M_2$ (no difference between M_1 and M_2)

Alternative hypothesis:

2. $H_1 = M_1 \neq M_2$
3. $H_1 = M_1 < M_2$
4. $H_1 = M_1 > M_2$

where M_1 = mean of the experimental group; and

M_2 = mean of control group.

When it is said that the mean of the experimental group will be higher than the mean of the control group, we are concerned with only one end of the distribution. Putting it in terms of a normal curve, we are concerned with only one end of the curve (see Figure 23.5). When the alpha level is set at the 0.05 level, we have a 5% of area of the normal curve all in one tail rather than having distributed it equally into two tails of the curve. Therefore, the directional null hypothesis is called a one-tailed test. A simple inspection of the table of areas of the normal curve given at the end of this chapter reveals that a *z* score of 1.64 cuts off 5% of the area under normal curve in the smaller part, and similarly a *z* score of 2.33 cuts off 1% of the area in the smaller part. If the null hypothesis is rejected, that is hypothesis 1 is not tenable, we automatically accept the alternative hypothesis. If the experimenter has somehow reason to believe that the experimental group would have a lower mean score than the control group (alternative hypothesis), he can set up a directional hypothesis that the mean of the experimental group is lower than the mean of the control group (one-tailed test). Rejection of this hypothesis would automatically lead to the acceptance of the above null hypothesis. This time the area of the normal curve in which the experimenter is interested is 5% or 1% of the area in the left-hand tail of the normal curve. When the null hypothesis is rejected by using a one-tailed test, we say that we are rejecting the hypothesis at 1% or 5% *points*, not levels.

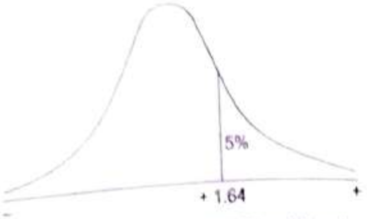

Fig. 23.5 One-tailed test at 0.05 or 5% point

A two-tailed test is one in which the investigator is interested in evaluating the difference between the groups. The direction of difference is of no importance here. The null hypothesis will be that the mean of the experimental group is equal to the mean of the control group, that is, there is no difference between the means of the experimental group and the control group. The alternative hypothesis would be that the mean of the experimental group is not equal to the mean of the control group. Thus, we show our concern with both tails of the distribution. At the 5% level, we have 5% of the area under the normal curve equally divided at both the tails (see Figure 23.6). From the table of areas of the normal curve it can easily be read that a z score of +1.96 cuts off 2.5% area at both extremes. Since a normal curve is bilaterally symmetrical, a z score of −1.96 would also cut off 2.5% area in the left-hand tail. In a two-tailed test, a negative z score is interpreted in the same way as a positive z score. When the null hypothesis is rejected by using a two-tailed test, it is said that we have rejected it at 5% or 1% *levels* and not at 5% or 1% points.

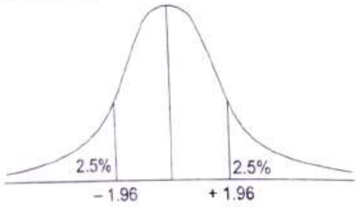

Fig. 23.6 Two-tailed test at 0.05 or 5% level

There are some problems with one-tailed test and a good researcher must take into account those problems. One important problem with one-tailed tests is that they allow the researcher to reject null hypothesis (H_0) even when the difference between the sample and the population is relatively small. In other words, one-tailed tests make it too easy to make a Type I error (rejecting a true null hypothesis) Therefore, most researchers consider one-tailed tests as improper. That is the reason why two-tailed tests are always more acceptable and generally preferred.

Another problem with one-tailed test arises from the fact that they look for a treatment effect in one direction only. Let us illustrate it with an example: Suppose the researcher wants to examine the effect of background music on productivity of the factory workers. The productivity is measured in terms of Mean performance of a period of 30 days. Further, suppose that the researcher develops the one-tailed test stating that background music will tend to enhance the mean performance of the factory workers. In this case since one-tailed test has no critical region on the left-hand side of mean (in normal probability curve), it would not detect a significant decrease in productivity. If the researcher had developed two-tailed tests, they would have been sensitive to a change in either direction. For this reason, most researchers don't like to use one-tailed test.

Power of test

As we have just considered, there is an inversed relation between the likelihood of committing Type I and Type II error. In other words, a decrease in Type I error will increase Type II error for given sample of *N* elements. *If the researchers wish to reduce both Type I and Type II errors, he must increase N.* Different types of statistical tests offer the possibility of different balances between these two types of errors. For obtaining such balance, notion of power of statistical test is very important. The power of a statistical test is defined as the probability of rejecting the null hypothesis when, in fact, it is false. In terms of equation, it can be stated like:

$$\text{Power} = 1 - \text{probability of Type II error}$$

The power of a statistical test generally increases with increase in size of N. It is clear from the above interpretation that the power of test and concept of Type II error is closely related. When a treatment effect exists, the hypothesis testing can have any one of the two conclusions:

(a) It can fail to identify treatment effect (a Type II error)

or,

(b) It can correctly identify the treatment effect (rejecting a false null hypothesis).

If, for example, a hypothesis testing has 20% chance of failing to identify the treatment effect, then it must have an 80% chance of correctly identifying it. That is the reason why the power of statistical test is determined by 1 – probability of Type II.

Researchers have identified three factors that tend to affect the power of a statistical test: alpha level, choice of one-tailed or two-tailed test and size of the sample. A discussion follows.

(a) *The alpha level:* Alpha level is the level of significance. When alpha level is reduced, the risk of Type I error (rejecting a true H_o) is also reduced by making it more difficult to reject H_o. Since a small alpha level generally reduces the probability of rejecting the null hypothesis, it also reduces the statistical power of the test.

(b) *One-tailed vs. Two-tailed tests:* As we know, one tailed tests make it very easier to reject the null hypothesis. Since one-tailed tests increase the probability of rejecting H_o, they also tend to increase the power of the test.

(c) *Sample size:* The power of statistical test is also affected by the size of the sample. As we know, when the size of the sample is larger, it tends to represent the population in a better way. If there is a real treatment effect in the population, a larger sample will be more likely to locate it than a smaller sample. Therefore, the power of test is enhanced by increasing the size of the sample.

Thus we find that the power of test is influenced by various factors.

Now, we are in a position to proceed ahead for showing the calculation of *t* ratio from different groups. Ordinarily, three types of situations arise while one is calculating the *t* ratio.

1. *t* ratio from independent groups
2. *t* ratio from correlated groups
3. *t* ratio from matched-groups

Regardless of the nature of the group, *t* ratio is calculated by the following equation:

$$t = \frac{(M_1 - M_2) - 0}{SE_D} \qquad (23.7)$$

where, M_1 = mean of the first group;
M_2 = mean of the second group;
SE_D = standard error of the difference between two sample means.

In each of the above three groups, the formula for SE_D varies and hence, we shall illustrate the calculation of the *t* ratio from each of the above types of groups separately.

t ratio from independent groups: Two groups are said to be independent when no correlation exists between them. Suppose one group of boys (N = 20) and one group of girls (N = 22) were administered a mechanical reasoning test. Their data were summarized as given on the next page.

$$N_1 = 20, \qquad N_2 = 22, \qquad M_1 = 34.56, \qquad M_2 = 30.56,$$

$$SD_1 = 5.68, \qquad SD_2 = 6.98$$

Do the two groups differ on the measure of mechanical reasoning test?

The above question can be answered by computing the *t* ratio. For independent groups, SE_D may be calculated with the help of the following equation:

$$SE_D = \sqrt{SE_{M_1}{}^2 + SE_{M_2}{}^2} \qquad (23.8)$$

where, SE_D = standard error of the difference between means;

SE_{M_1} = standard error of the first mean; and

SE_{M_2} = standard error of the second mean.

Here standard error of mean can be calculated as given below:

$$SE_M = \frac{SD}{\sqrt{N-1}} \qquad (23.9)$$

Substituting the value in Equation 23.9, we have

$$SE_{M_1} = \frac{5.68}{\sqrt{19}} = 1.302; \quad SE_{M_2} = \frac{6.98}{\sqrt{21}} = 1.524$$

Now, SE_D becomes equal to:

$$SE_D = \sqrt{(1.302)^2 + (1.524)^2} = 2.004$$

$$\therefore \quad t = \frac{(M_1 - M_2) - 0}{SE_D} = \frac{(34.56 - 30.56) - 0}{2.004} = \frac{4.00}{2.004} = 1.996$$

$$df = (N_1 - 1) + (N_2 - 1) = (20 - 1) + (22 - 1) = 40$$

Entering the probability table of *t* ratios, we find that the obtained *t* at $df = 40$ is significant at the 0.05 level but not at the 0.01 level. Hence, we conclude that the null hypothesis is rejected and that there is a true difference between the means of the groups of boys and girls.

t ratio from correlated groups: The correlated groups are those which exhibit some correlation with each other on the given measures. One of the most convenient way of getting the correlated means is to repeat the same test on the same group twice on two different occasions. In between the initial and the final administration of the test, some experimental treatment is given to the group. This is known as the *repeated measure t test.* Because the group and the test are the same, it is highly probable that there will be a correlation between the initial measures and the final measures. Suppose a group of 20 students of Class IV is administered an English spelling test in January. The obtained mean score and standard deviations are given on the next page. After a year's training in spelling, they are again given the same test. This time their mean is considerably raised and the standard deviation is lowered. The correlation coefficient between the initial and final set of scores was positive and significant.

Does training produce a significant difference between the intial mean and the final mean?

	Initial test	Final test
N_S	20	20
Means	36.28	40.33
SD_S	4.28	2.32
Coefficient of correlation	0.80	

In case of testing the significance of mean difference between two correlated means, SE_D is calculated with help of the following formula:

$$SE_D = \sqrt{SE_{M_1}{}^2 + SE_{M_2}{}^2 - 2r_{12}SE_{M_1}SE_{M_2}} \qquad (23.10)$$

where r_{12} = coefficient of correlation between the initial set of scores and the final set of scores. The rest of the subscripts are defined like those in Equation 23.8.

The standard error of mean (SE_M) is calculated by Equation 23.9. Thus,

$$SE_{M_1} = \frac{4.28}{\sqrt{19}} = 0.982, \qquad SE_{M_2} = \frac{2.32}{\sqrt{19}} = 0.532$$

$$SE_D = \sqrt{SE_{M_1}{}^2 + SE_{M_2}{}^2 - 2r_{12}SE_{M_1}SE_{M_2}}$$

$$= \sqrt{(0.982)^2 + (0.532)^2 - 2(0.80)(0.982)(0.532)}$$

$$= \sqrt{0.9643 + 0.2830 - 0.8358} = 0.641$$

$$t = \frac{(M_1 - M_2) - 0}{SE_D} = \frac{(36.28 - 40.33) - 0}{0.625} = \frac{4.05}{0.641} = 6.31$$

$$df = N - 1 = 20 - 1 = 19$$

Entering the probability table of *t* ratios at $df = 19$, we find that our obtained value of *t* exceeds the value of *t* at even the 0.001 level of significance. Hence, the null hypothesis is rejected and it is concluded that the training has produced significant difference between the mean of the initial set of scores and the mean of the final set of scores.

t ratio from matched-groups: Sometimes it becomes necessary for the researchers to match the groups under study. Matching can be done on the basis of numbers or it can be done in terms of mean and standard deviation. When the matching is done on the basis of the number of the subject, each person has his corresponding match in the other group and therefore, the number of persons in the two matched groups is *always* equal. When matching is done in terms of mean and standard deviation, the number in the two groups may or may not be equal.

Suppose we take two groups from two different classes and each group is compared on the basis of numerical reasoning test. Both groups have been matched in terms of mean and standard deviations on the basis of scores on general intelligence Test. Do the groups differ in terms of mean numerical ability? The data are given on the next page.

	Class VIII	Class X
N_1	100	120
Means of Intelligence Tests	70.26	70.25
SDs on Intelligence Tests	9.98	10.02
Means on Numerical Reasoning Tests	55.62	60.34
SDs on Numerical Reasoning Tests	8.67	7.58
Coefficient of Correlation between General Intelligence Test scores and Numerical Reasoning Test scores	0.45	

The equation for calculating SE_D in case of groups matched in terms of means and SD is given below.

$$SE_D = \sqrt{\left(SE_{M_1}{}^2 + SE_{M_2}{}^2\right)(1-r^2)} \tag{23.11}$$

Where, subscripts are defined as usual. Now, we can proceed as under:

$$SE_{M_1} = \frac{8.67}{\sqrt{100}} = 0.867; SE_{M_2} = \frac{7.58}{\sqrt{120}} = 0.691$$

$$SE_D = \sqrt{[(0.867)^2 + (0.691)^2]\,(1-0.45^2)}$$

$$= \sqrt{(0.7516)+(0.4777)(1-0.2025)}$$

$$= \sqrt{(1.2293)(0.7975)} = 0.991$$

$$t = \frac{(M_1-M_2)-0}{SE_D} = \frac{(55.62-60.34)-0}{0.991}$$

$$= \frac{4.72}{0.991} = 4.762$$

$$df = (N_1-1)+(N_2-1)-1 = (100-1)+(120-1)-1$$

$$= (99+119)-1 = 218-1 = 217$$

Entering the probability table of *t* ratios at *df* 217, we find by interpolation that the obtained value of *t* exceeds the *t* value required at even the 0.001 level of significance. Rejecting the null hypothesis, we conclude that the two groups differed significantly in terms of numerical mean ability.

2. Analysis of Variance: *F* Ratio

The *t* ratio or *z* ratio is one of the powerful parametric tests through which we can test the significance of the difference between two means. There are two general limitations of the *t* ratio. First, when there are several groups and if we want to test the significance of the mean difference among them, several *t* ratios are required to be computed. For example, suppose there are five groups. Then we need to compute

$$\frac{N(N-1)}{2} = \frac{5\times 4}{2} = 10\ t \text{ ratios}$$

which is, of course, a cumbersome job. Second, the *t* ratio does not account for interaction effect in its statistical analysis. The variations in the scores may be due to the interactions taking place among groups. Such variations are not accounted for by *t* ratios. In order to remove these two limitations we turn to *analysis of variance*, originally developed by R A Fisher. Analysis of variance is a class of statistical techniques through which we test the overall difference among the two or more than two (normally more than two) sample means. Analysis of variance is of two common types: *Simple analysis of variance* or *one-way analysis of variance* and *complex analysis of variance* or *two-way analysis of variance*. Analysis of variance (of whatever type) is often referred to by its contraction, ANOVA.

Before we take up discussion on ANOVA, a fundamental concept of ANOVA, that is, the *concept of sum of squares* must be properly understood. Let us take an example to illustrate the meaning of the concepts of sum of squares. Suppose we have conducted a study with two groups of students concerning their attitude towards legalization of a (banned) drug. Each member of group A of students had never taken that drug and each member of the group B had taken this drug at least once in a week. A high score on the attitude scale indicated that students strongly favoured legalization. Here we want to test the hypothesis that there is no difference in this attitude towards legalization because two groups of students have been taken from the same population. Table 23.3 presents the scores of Group A and Group B towards the legalization of drug.

Table 23.3 Group A and Group B attitude towards the legalization of drug

Group A		Group B	
X_1	X_1^2	X_2	X_2^2
3	9	8	64
2	4	10	100
3	9	12	144
8	64	10	100
Sum: 16	86	40	408

$N_1 = 4,$ $\qquad$ $N_2 = 4.$

$$\overline{X}_1 = \frac{16}{4} = 4, \qquad \overline{X}_2 = \frac{40}{4} = 10.$$

$$\Sigma X_{\text{total}} = 16 + 40 = 56$$

$$\overline{X}_{\text{total}} = \frac{56}{8} = 7$$

In Table 23.3, the mean for Group A is 4 and the mean for Group B is 10. The overall mean is 7. For obtaining total sum of squares (SS_{total}), it is required that from each individual score the overall mean must be subtracted, squared and added. Thus

$$SS_{\text{total}} = \Sigma(X - \overline{X}_{\text{total}})^2 \tag{23.12}$$

For the data in Table 23.3, the total sum of squares will be then

$$SS_{\text{total}} = (3-7)^2+(2-7)^2+(3-7)^2+(8-7)^2+(8-7)^2+(10-7)^2+(12-7)^2+(10-7)^2$$

$$= 16+25+16+1+1+9+25+9 = 102$$

In this way, we find that the total sum of squares (SS_{total}) is the sum of squared deviations from overall mean and therefore, is a measure of total variance of the distribution.

Another way of calculating total sum of squares is through *raw score equation* as under:

$$SS_{\text{total}} = \Sigma X^2_{\text{total}} - \frac{(\Sigma X_{\text{total}})^2}{N} \tag{23.13}$$

$$= 86 + 408 - \frac{(16+40)^2}{8} = 494 - 392 = 102$$

In analysis of variance (ANOVA), the total sum of squares is divided into two parts: within group sum of squares (SS_w) and between group *sum of squares* (SS_b). The within sum of squares component is the measure of variability of the dependent variable scores within each category or group of independent variable and therefore, is the variation in the dependent variable that can't be attributed to the independent or treatment variable. The source of the within-group variability includes all the variables influencing dependent variables other than the independent or treatment variable. Thus all within-group variability must come from variables other than the *specified independent variables.* The within group sum of squares, in the above example, measures the sum of the squared deviations of scores in Group A around $\overline{X}_1$ plus the sum of the squared deviations of scores in Group B around $\overline{X}_2$.

For the data in Table 23.3 the within group sum of squares may be calculated as under:

$$\Sigma(X_1 - \overline{X}_1)^2 + \Sigma(X_2 - \overline{X}_2^2)^2 \tag{23.14}$$

$$= (3-4)^2 + (2-4)^2 + (3-4)^2 + (8-4)^2 + (8-10)^2 + (10-10)^2 + (12-10)^2 + (10-10)^2$$

$$= 1 + 4 + 1 + 16 + 4 + 0 + 4 + 0 = 30$$

Now, raw score equation for within group sum of squares may also be used to arrive at the same result, that is, 30.

$$SS_w = SS_1 + SS_2 \tag{23.15}$$

$$SS_1 = \Sigma X_1^2 - \frac{(\Sigma X_1)^2}{N_1} \tag{23.16}$$

$$= 86 - \frac{(16)^2}{4} = 86 - 64 = 22$$

$$SS_2 = \Sigma X_2^2 - \frac{(\Sigma X_2)^2}{N_2}$$

$$= 408 - \frac{(40)^2}{4} = 408 - 400 = 8$$

$$SS_w = 22 + 8 = 30$$

The *between sum of squares* (SS_b) is the sum of the sum of squares between each group and directly reflects the impact of the independent variable or treatment variable. In other words, the between sum of squares component is a measure of the variation between the groups of the independent variable and therefore, is the variation in the dependent variable that is attributable to the independent or treatment variable. It may be obtained by subtracting the overall mean from each group mean, squaring the result, multiplying by N in each and finally summing across all the groups. Thus

$$SS_b = \Sigma N(\overline{X}_i - \overline{X}_{\text{total}})^2 \tag{23.17}$$

where, N_i is the number in the ith group and $\overline{X}_i$ is the mean of ith group.

$$SS_b = 4(4-7)^2 + 4(10-7)^2 = 36 + 36 = 72$$

The raw score formula for calculating the between-group sum of squares is:

$$SS_b = \Sigma \frac{(\Sigma X_i)^2}{N_i} - \frac{(\Sigma X_{\text{total}})^2}{N} \tag{23.18}$$

$$= \left\{\frac{(16)^2}{4} + \frac{(40)^2}{4}\right\} - \frac{(56)^2}{8} = (64 + 400) - 392 = 464 - 392 = 72$$

However, most simply SS_b can be calculated by subtracting the value of SS_W from SS_{tot}:

$$SS_b = SS_{\text{total}} - SS_w \tag{23.19}$$

$$= 102 - 30 = 72$$

In simple analysis of variance, there is only one independent variable and the samples are classified into several groups on the basis of this variable. Since the basis of classification is only one independent variable, the simple analysis of varaince is also known as one-way ANOVA. Such ANOVA is suited to the *completely randomized design*. In complex ANOVA there are two or more than two independent variables, which form the basis of classification of groups. Such ANOVA is suited to *factorial design*.

Statistically, the F ratio is calculated as follows.

$$F = \frac{\text{Larger variance}}{\text{Smaller variance}} \quad \text{or} \quad \frac{\text{Between-groups variance}}{\text{Within-groups variance}}$$

Between-groups variance refers to variation in the mean of each group from the total or grand mean of all groups. Within-groups variance refers to the average variability of scores within each group. The theme of the analysis of variance is that if the groups have been randomly selected from the population, these two variances, namely, between-groups variance and within-group variance are the unbiased estimates of same population variance. The significance of difference between these two types of variances is tested through the F test.

Anova has some assumptions which should be met.

(a) The population distribution for each treatment condition should be normal. In other words, there should be normality within groups who have been measured on dependent variable.

(b) The individuals who have been observed should be distributed randomly in the groups.

(c) The dependent variable should be measured on interval scale and independent variable (s) should be measured on nominal scale (Elifson 1990)

(d) Within-groups variances must be approximately equal or homogeneous. This is referred to as *homogeneity of variances*. There are ordinarily two popular methods of determining the homogeneity of variances: Bartlett's test of homogeneity of variances and Hartley's F-Max test. For Bartlett's test, readers are referred to consult Kirk's (1982) *Experimental Design*. Here we shall concentrate upon Hartley's F-Max test only.

Harley's F-Max test is based on the principle that a sample variance presents an unbiased estimate of the population variance. Therefore, if the population variances are the same, the sample variances should be ordinarily very similar. The procedure for computing F-Max test is as under:

(a) Complete sample variance for each of the separate samples. The sample variance is computed by $\frac{SS}{n-1} = \frac{SS}{df}$ where SS is the sum of squared deviations of each sample separately and calculated by $\Sigma X^2 - \frac{(\Sigma X)^2}{n}$.

(b) Select the largest and the smallest of these sample variances and compute F-max as under:

$$\text{F-max} = \frac{\text{Langese sample variance}}{\text{Smallese sample variance}} \qquad (23.20)$$

A relatively large value of F-Max test indicates that there exists a large difference among sample variances. In such a situation, the data suggest that the population variances are different and that the assumption of homogeneity of variances has been violated. On the other hand, the small value of F-Max test (that is, near 1.00) indicates that the sample variances are similar and therefore, the assumption of homogeneity of variance is reasonable. Let us illustrate it with an example.

Suppose there are three independent samples, each having $n = 10$. The sample variances are 12.66, 10.78, 11.79. For these data,

$$\text{F-Max} = \frac{\text{Largest sample variance}}{\text{Smallest sample variance}} = \frac{12.66}{10.78} = 1.17$$

The data do not provide evidence that the assumption of homogeneity of variance has been violated because the value of F-Max test is near 1.

There is also another way of reaching a at decision about the assumption of homogeneity of variance by comparing the value of F-Max test with the critical value provided in Table[*]. For using this Table, we need to know k = number of separate samples, $df = n - 1$ and alpha level as predetermined by the investigator. For the above data, we have $k = 3$. $df = n - 1 = 10 - 1 = 9$ and a value of 5.34 at the alpha level of 0.05 and a value of 8.5 at the alpha level of 0.01 is needed. Since the obtained value of F-Max test is less than 5.34, we may conclude that the assumption of homogeneity of variance is reasonable. Had the obtained value been larger than the value given in the Table, we would have concluded that the homogeneity assumption is not valid.

For an illustration of simple analysis of variance, with independent measures look at Table 23.4 in which the scores of the three groups of subjects A, B and C are shown on the educational achievement test, and ANOVA has been calculated from those scores.

Grand sum (ΣX): $561 + 154 + 312 = 1027$

Grand sum of squares (ΣX^2): $38611 + 2610 + 16070 = 57291$

Step 1: Correction (C): $\dfrac{(\Sigma X)^2}{N} = \dfrac{(1027)^2}{30} = 35157.63$

Step 2: Total sum of squares (TSS) : $\Sigma X^2 - C$

$$= 57291 - 35157.63$$
$$= 22133.37$$

---

[*] For this table, the reader is referred to the book entitled *Statistics for Behavioural Sciences* by F J Gravetter & LB Wallnau (1987).

Table 23.4 Simple ANOVA based on the hypothetical scores of three groups

Gr. A X	Gr. B X	Gr. C X	Gr. A X^2	Gr. B X^2	Gr. C X^2
78	10	50	6084	100	2500
58	14	58	3364	196	3364
40	9	65	1600	81	4225
30	12	72	900	144	5184
10	20	10	100	400	100
65	25	9	4225	625	81
88	12	8	7744	144	64
87	18	10	7569	324	100
80	14	16	6400	196	256
25	20	14	625	400	196
Sum: 561	154	312	38611	2610	16070

Step 3: Between (or among) sum of squares (BSS):

$$\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2^2)}{n_2} \ldots\ldots + \frac{(\Sigma X_n)^2}{n_n} - C$$

$$= \frac{(561)^2}{10} + \frac{(154)^2}{10} + \frac{(312)^2}{10} - 35157.63$$

$$= \frac{314721 + 23716 + 97344}{10} - 35157.63$$

$$= 43578.10 - 35157.63 = 8420.47$$

Step 4: Within sum of squares: TSS – BSS $= 22133.37 - 8420.47 = 13712.90$

Summary: Analysis of Variance

Source of variation	*df*	Sum of squares	Mean square or variance
Between-groups	$k - 1 = 3 - 1 = 2$	8420.47	4210.235
Within-groups	$N - k = 30 - 3 = 27$	13712.90	507.885
Total	$N - 1 = 30 - 1 = 29$	22133.37	

$$F = \frac{\text{Between-groups variance}}{\text{Within-groups variance}} = \frac{4210.235}{507.885} = 8.289 = 8.29$$

Since there are 30 cases in the above problem, we have $N - 1 = 30 - 1 = 29df$ in all. *df* for between-groups is equal to the number of groups (*K*) minus one. Since there are three groups, hence *df* for between-groups is $K - 1 = 3 - 1 = 2$, *df* for within-groups is equal to the total number of cases minus number of groups (*K*). Hence, it is equal to $N - K = 30 - 3 = 27$. After calculating the number of degrees of freedom and the sum of squares for each of the three sources of variations, we compute mean squares or variances, which are obtained by dividing each of the sum of squares by its respective number of degrees of freedom. These two types of variances are the estimates of the population variance. We obtain *F* ratio by dividing the between-groups variance by the within-groups variance.

F ratio is interpreted by the use of the *F* Table (Guilford & Fruchter 1978). In the *F* Table, the number of degrees of freedom for greater mean square (df_1) is written at the top and the number of degrees for freedom for smaller mean square (df_2) is written on the left-hand side. For this problem, $df_1 = 2$ and $df_2 = 27$. Locating at these *dfs*, we find that the required *F* ratio at the 0.05 level is 3.35 and at the 0.01 level is 5.49. Since the obtained value of *F* ratio is 8.29, which exceeds 5.49, we reject the null hypothesis and conclude that there is an overall difference between the three groups of subjects on the educational achievement test.

Tests after the F test (Post Hoc tests)

One general limitation of the *F* test is that it only tells about the overall difference between the groups under study but tells nothing about the location of the exact difference. *For example*, in the above problem the obtained *F* ratio of 8.29 is significant, which definitely indicates that there is significant difference between the groups under study, but whether the significant difference is between A and B or A and C or B and C cannot be said. Therefore, when *F* ratio is significant, we need some additional tests after this *F* test.

In fact, in ANOVA the null hypothesis states that there is no treatment effect and therefore, all sample means are the same. Although this appears to be a very simple conclusion, in most cases, it creates many problems. For example, suppose there are only two treatment groups in the experiment, null hypothesis will state that the mean of one treatment group does not differ from the mean of the other treatment group. If this hypothesis is rejected, the straightforward conclusion is that the two means are not equal. But when there are three (or more) treatment groups, the problem gets complicated. With $k = 3$, for example, rejecting null hypothesis indicates that not all means of the treatment group are the same. Now here, the researcher has to decide which ones are different. Is Mean_1 is different from Mean_2? Is Mean_2 different from Mean_3? Is Mean_1 is different from Mean_3? Are all three different? The purpose of *post hoc tests* is to answer these questions.

Post hoc tests are done after an analysis of variance has been carried out. These are called *post hoc comparisons* because they are done *after the fact* and not planned in advance (Aron, Aron & Coups 2006). In general, a post hoc test enables the researcher to go back through the data and compare the two individual treatments at a time, that is, in pair. In statistical terms, this is called making *pairwise comparison*.

Although there are different post hoc test procedures because there are many different ways of controlling the alpha levels, these tests can be classified into two broad categories: *a priori or planned comparisons* and *a posteriori or unplanned comparisons*.

A priori tests are meant to compare the specific treatment conditions that are identified before the experiment begins. Such tests generally make little effort to control the alpha level. Planned comparisons can be made even when the overall *F* ratio is not significant. The rationale behind the tests is that the researcher has a specific small experiment (comparing two treatments), which is contained within a larger experiment. Since the research has specific hypotheses for just this portion of experiment, it is reasonable to test its results separately.

A posteriori tests are used for treatment comparisons that were not necessarily planned at the start of the experiment. In fact these tests try to control the overall alpha level by making adjustments for the number of different potential comparisons in the experiment. A posteriori test is justified when *F* ratio from over all ANOVA is significant. If the overall *F* is not significant, comparison can still be made to simply check the differences between means. It may, however, be noted that *a priori* comparisons, here, are more powerful than *post hoc* comparison. If the overall *F* is significant and the researcher is using *post hoc test*, he should not use the conventional *t* test because the alpha level at which he would be testing will be numerically higher than the alpha level originally specified. This is likely to happen because some of the obtained significant differences may be due to chance factors. It may also be clearly noted that as the number of comparisons increases, the probability of Type I error also increases.

A variety of post hoc tests are available. For a given situation, different tests may be used and the outcomes of these different tests may be different. Statisticians who are working with these post hoc comparisons are themselves not in complete agreement regarding the appropriateness of the test. Important post hoc tests are: *Newman-Keuls test, Duncan's Multiple Range test, Protected t test*, Tukey's *Honestly Significant Difference* (HSD) *test and Scheffe test*. Of these, the two most popular post hoc tests are Tukey's *Honestly Significant Difference* (HSD) *test* and *Scheffe test* which will be discussed. Tukey HSD test allows the researcher to compute a single value that determines the minimum difference between treatment means that is considered essential for significance. This value is called as the Honestly Significant Difference or HSD, which is then, used to compare any two treatment conditions. If the mean difference exceeds Tukey's HSD, the researcher concludes that there is a significant difference between treatments. If the mean difference is less than Tukey's HSD, he concludes that there is no significant difference between the treatments. The formula for Tukey's HSD is as under:

$$HSD = q\sqrt{\frac{MS_{within}}{n}} \tag{23.21}$$

where, MS_{within} = within treatment variance

n = number of scores in each treatment

q = studentized range statistic.

For further details of Table of Studentized Range statistic, readers are referred to consult Gravetter and Wallnau (1987, A-35).

Taking example from Table 23.4, we can calculate HSD for making three post hoc comparisons and determining in which pair(s) the significant mean difference actually lies.

From Table 23.4,

$MS_{within} = 507.885$

$n = 10$

K (or number of treatments/columns) = 03

Alpha level for q=.05, df for error term = $N - k = 30 - 3 = 27$

$\therefore \quad q = 3.51$ (by interpolation from Table 23.5)

$$HSD = 3.51\sqrt{\frac{507.885}{10}}$$

$$= 3.51\sqrt{50.79} = 3.51 \times 7.13 = 25.02$$

Using this value (25.02), we can make the following conclusions:

1. Mean difference between A and B = 56.10 – 15.4 = 40.7 (Significant)
2. Mean difference between A and C = 56.10 – 31.2 = 24.9 (Not Significant)
3. Mean difference between C and B = 31.2 – 15.4 = 15.8 (Not Significant)

Since mean difference between A and B only exceeds 25.02, it is regarded as significant. The other two mean differences, that is, between A and C and B and C are less than 25.02, so they are not significant. In this example, the mean difference between any two samples must be at least 25.02 to be significant.

Table 23.5 A selected portion of the value of studentized range statistics (q)*

(Light face = 0.5; **Bold face = 0.01**)

	K = Number of treatments					
df for error term	2	3	4	5	6	7
8	3.26	4.04	4.53	4.89	5.71	5.40
	4.75	**5.64**	**6.20**	**6.62**	**6.96**	**7.24**
24	2.92	3.53	3.90	4.17	4.37	4.54
	3.96	**4.55**	**4.91**	**5.17**	**5.37**	**5.54**
30	2.89	3.49	3.85	4.10	4.30	4.46
	3.89	**4.45**	**4.80**	**5.05**	**5.24**	**5.40**
∞	2.77	3.31	3.63	3.86	4.03	4.17
	3.64	**4.12**	**4.40**	**4.60**	**4.76**	**4.88**

Following the Scheffe technique also, we can locate the difference between three means. Since there are three groups, three comparisons are likely to be made, namely, A vs B, A vs C and B vs C. Following the Scheffe technique, we are required to compute the *F* ratio with the help of Equation 23.22 for each of three groups.

$$F = \frac{(M_1 - M_2)^2}{SD_w^2(N_1 + N_2)/N_1N_2} \qquad (23.22)$$

Now, three *F*s for three pairs of distribution can be calculated as follows:

F ratio for distributions A and B:

$$F = \frac{(56.10 - 15.40)^2}{507.885\,\frac{(10+10)}{(10)(10)}} = 16.31$$

F ratio for distributions A and C:

$$\frac{(56.10 - 31.20)^2}{507.885\,\frac{(10+10)}{(10)(10)}} = 6.10$$

---

* Abstracted from Table B.5 of *Statistics for Behavioural Sciences* by F S Gravetter & L A Wallnau (1987).

F ratio for distributions B and C:

$$\frac{(15.40 - 31.20)^2}{507.885\,\frac{(10+10)}{(10)(10)}} = 2.46$$

As pointed out earlier, *F* at the 0.05 level of significance for $df_1 = 2$ and $df_2 = 27$ is 3.35. This value, if multiplied by $k - 1$, yields $(3 - 1)(3.35) = 6.70$. Only the *F* ratio for distributions A and B is greater than 6.70. Hence, it is concluded that there is a significant difference between the means of A and B only. The mean difference between A and C and B and C is not significant. The same conclusion had been arrived by using Tukey's Honestly Significant Difference Test.

So far as the computation of the two-way analysis of variance is concerned, the reader is referred to Chapter 21 where details of statistical calculations have been shown.

Analysis of variance (ANOVA) may be univariate analysis or multivariate analysis. In univariate analysis there is one dependent variable and there may be more than one independent variable. In multivariate analysis of variance, there are more than one dependent variable. Usually dependent variables are different measures of approximately the same thing such as two different reading ability tests. This is called as multivariate analysis of variance (MANOVA) which is different from an ordinary analysis of variance because in it (MANOVA), there is more than one dependent variable. When the researcher finds an overall significant difference among groups with MANOVA, this means that the groups differ on combination of dependent variables. Subsequently, the researcher proceeds to know whether the groups differ on any or all of the dependent variables considered individually. In this way, MANOVA is followed by an ordinary analysis of variance for each of the dependent variables. In the present text, no attempt will be made to illustrate the computation of MANOVA.

Likewise, the analysis of covariance in which there are more than one dependent variable is called as multiple analysis of covariance (MANCOVA). MANCOVA differs from an ordinary analysis of covariance (ANCOVA) because in the former, there are more than one dependent variable where as in the latter, there is only one dependent variable. In the present text, MANCOVA will not be illustrated with numerical examples. However, ANCOVA will be discussed in detail with numerical examples.

3. Analysis of Covariance (ANCOVA)

The analysis of covariance (ANCOVA) was developed by R A Fisher and the very first example of its use appeared in literature in 1932. Most of these applications are from agricultural experimentations. A close examination of these examples in agriculture also helps to clarify its application in the behavioural sciences. Analysis of covariance, a most widely used elaboration of analysis of variance, is a technique in which indirect method or statistical control is employed to enhance the precision of the experiment. In its procedure or methodology, ANCOVA may be equated with partial correlation where the researcher seeks a measure of correlation between two sets of variables—dependent and independent—by partialling out the impact of third or intervening variables. In ANCOVA the researcher also tries to partial out the side effects, if any, in the experiment due to lack of exercising proper experimental control over the intervening variables or covariates (or covariables). Each of the variables controlled for (or partialed out or held constant) is called as *covariate*. In ANCOVA, the statistical control is achieved by including measures on concomitant variate (*X*). This is uncontrolled variable, also called as covariate and is not itself of experimental interest. The other variate which is of experimental interest is termed as the criterion and designated as *Y*. In this way in ANCOVA, the researcher obtains two observations (*X* and *Y*) from each participant. Measurements on *X* (covariate) obtained before giving treatment are made to adjust the measurements on *Y* (criterion). When *X* and *Y* are associated, a part of the variance of *Y* occurs due to the variation in *X*. In fact, ANCOVA is a method of making adjustments that are linked to the problem of correlation. Here the problem is to specify how much of variation of *Y* can be predicted from variation in *X* and then subtracting this to obtain the remaining (or leftover) variation as an adjusted value. A formula for the purpose of correcting Y scores for initial differences in X scores is as under:

$$SS_{y.x} = SS_y - (SP)^2/SS_x \tag{23.23}$$

where, $SS_{y.x}$ = sum of squares of Y when variability contributed by X has been controlled or removed.

SS_y = sum of squares of Y scores

SP = sum of products of deviations of X and Y

SS_x = sum of squares of X scores

In a nutshell, it can be said that ANCOVA is a method of analysis that enables the investigator to equate pre-experimental status of the groups in terms of relevant known variables. Differences in the initial status of the groups are removed statistically so that they can be compared as though their initial status had been equated. The scores that have been corrected by this procedure are technically known as *residuals* because they are what remains after the inequalities have been corrected or removed. Results are interpreted like any other analysis of variance with one major exception. While reporting results, instead of giving the means of each group, the researcher may give the *adjusted means*, the means of each group after adjusting or partialing out the effect of the covariates.

The assumptions of ANCOVA are as under:

(i) The X scores (covariate) are not affected by treatments.

(ii) The dependent variable that is under measurement should be normally distributed in the population.

(iii) The regression coefficient of the final scores (Y) on initial scores (X) should be more or less same in all groups and is linear, that is, correlations between X and Y scores are same for all groups.

(iv) Treatment groups are selected at random from the same population.

(v) Within-group variances should be approximately equal.

(vi) The contribution of variances in the total sample should be additive.

Numerical Example I

Three groups of five students each were randomly selected from class X of a school. They were, then, rated for their leadership quality and their scores were obtained. Subsequently, they were subjected to three different training techniques for improving their leadership qualities for a month. After a month of training, they were again assessed for their leadership qualities and their final scores were also obtained. The data so collected are presented in Table 23.6.

Table 23.6 Measures on a covariate (X) and a criterion variate (Y) for single factor experiment

Group A		Group B		Group C	
Initial scores X_1	Final scores Y_1	Initial scores X_2	Final scores Y_2	Initial scores X_3	Final scores Y_3
3	6	8	10	7	14
4	8	8	8	4	4
5	5	5	12	5	12
3	6	9	10	6	13
2	4	5	13	8	10
Sum: 17	29	35	53	30	53
Mean : 3.4	5.8	7.0	10.6	6.0	10.6

The above data calls for using ANCOVA. There may be many observable differences among the initial scores of the three groups because no attempt was made by the researcher to make these three groups equivalent at the start of the study, that is, before subjecting these groups to three different treatments. In the absence of such an experimental control, the researcher is forced to exercise statistical control by using ANCOVA. The computational work for ANCOVA may be started by arranging the above data as under (*cf.* Table 23.7).

Table 23.7 Computational arrangement for ANCOVA

Group A				
X_1	Y_1	$(X_1)(Y_1)$	X_1^2	Y_1^2
3	6	18	9	36
4	8	32	16	64
5	5	25	25	25
3	6	18	9	36
2	4	8	4	16
Sum: 17	29	101	63	177
Mean: 3.4	5.8			

Group B				
X_2	Y_2	$(X_2)(Y_2)$	X_2^2	Y_2^2
8	10	80	64	100
8	8	64	64	64
5	12	60	25	144
9	10	90	81	100
5	13	65	25	169
35	53	359	259	577
7.0	10.6			

Group C				
X_3	Y_3	$(X_3)(Y_3)$	X_3^2	Y_3^2
7	14	98	49	196
4	4	16	16	16
5	12	60	25	144
6	13	78	36	169
8	10	80	64	100
Sum: 30	53	332	190	625
Mean: 6	10.6			

$\Sigma X = 17 + 35 + 30 = 82$

$\Sigma Y = 29 + 53 + 53 = 135$

Steps of computation

A. ANOVA of X scores:

(i) Correction $(C_x) = \dfrac{(\Sigma X)^2}{N} = \dfrac{(17+35+30)^2}{15} = \dfrac{(82)^2}{15} = 448.27$

(ii) Total sum of squares $(SS_{xt}) = \Sigma X^2 - C_x$

$$= (63 + 259 + 190) - 448.27 = 512 - 448.27 = 63.73$$

(iii) Between group sum of squares $(SS_{xb}) = \dfrac{(\Sigma X_1)^2}{N_1} + \dfrac{(\Sigma X_2)^2}{N_2} + \dfrac{(\Sigma X_3)^2}{N_3} - C_x$

$$= \left(\frac{17^2}{5} + \frac{35^2}{5} + \frac{30^2}{5}\right) - 448.27$$

$$= (57.8 + 245 + 180) - 448.7 = 482.8 - 448.27 = 34.53$$

(iv) Within group sum of squares (SS_{xw}) = Total SS_x − Between SS_x

$$= 63.73 - 34.53 = 29.2$$

Table 23.8 Summary of ANOVA of *X* scores

Source of variation	SS	df	MS	F
Between groups	34.53	$k-1=3-1=2$	17.26	
Within groups (error)	29.2	$N-k=15-3=12$	2.43	7.10
Total	63.73	14		

$$F_x=\frac{17.26}{2.43}=7.10$$

F at 0.05 3.88 df 2,12
F at 0.01 6.93

B. ANOVA of *Y* scores:

(i) Correction $(C_y)=\frac{(\Sigma Y)^2}{N}=\frac{(29+53+53)}{15}=\frac{(135)^2}{15}=1215$

(ii) Total sum of squares $(SS_{yt})=\Sigma Y_2-C=(177+577+625)-1215=1379-1215=164$

(iii) Between groups sum of squares $(SS_{yb})=\frac{\Sigma Y_1^2}{N}+\frac{\Sigma Y_2^2}{N}+\frac{\Sigma Y_3^2}{N}-C_y$

$$=\frac{29^2}{5}+\frac{53^2}{5}+\frac{53^2}{5}-1215$$

$$=(168.2+561.8+561.8)-1215$$

$$=1291.8-1215=76.8$$

(iv) Within group sum of squares $(SS_{yw})=\text{Total } SS_y-\text{Between } SS_y$

$$=164-76.8=87.2$$

Table 23.9 Summary of ANOVA of *Y* scores

Source of variation	SS	df	MS	F
Between groups	76.8	$K-1=3-1=2$	38.4	
Within groups (error)	87.2	$N-k=15-3=12$	7.26	5.28
Total	114.94	14		

$$F_y=\frac{38.4}{7.26}=5.29$$

F at 0.05 3.88 df 2,12
F at 0.01 6.93

C. Analysis of covariance:

(i) Total sum of products $(SP_t)=(\Sigma X_1Y_1+\Sigma X_2Y_2+\Sigma X_3Y_3)-\frac{\Sigma X\cdot\Sigma Y}{N}$

$$=(101+359+332)-\frac{82\times135}{15}$$

$$=792-738=54$$

(ii) Sum of products between groups (SP_b)

$$=\left(\frac{\Sigma X_1\cdot\Sigma Y_1}{N_1}\right)+\left(\frac{\Sigma X_2\cdot\Sigma Y_2}{N_2}\right)+\left(\frac{\Sigma X_3\cdot\Sigma Y_3}{N_3}\right)-\left(\frac{\Sigma X\cdot\Sigma Y}{N}\right)$$

$$=\frac{(17\times29)}{5}+\frac{(35\times53)}{5}+\frac{(30\times53)}{5}-\frac{82\times135}{5}$$

$$=98.6+371+318-738$$

$$=787.6-738=49.6$$

(iii) Sum of products within groups $(SP_w)=SP_t-SP_b$

$$=54-49.6=4.4$$

For convenience, the sum of products (SP_t) may be arranged with sum of squares for both *X* and *Y* variates as under:

Table 23.10 Summary of sum of squares of X and Y as well as that of sum of products.

	Total	Within	Between
Sum of products (XY)	54	4.4	49.6
Sum of squares (X)	63.73	29.2	34.53
Sum of squares (Y)	164	87.2	76.8

(iv) Adjusted sum of squares for Y $(SS_{y.x})=SS_{yt}-\frac{(SP_t)^2}{SS_{xt}}$

$$=164-\frac{(54)^2}{63.73}=164-45.75=118.25$$

(v) Within group adjusted sum of squares $(SS_{y.x_b})$

$$=SS_{yw}-\frac{(SP_w)^2}{SS_{xw}}$$

$$=87.2-\frac{(4.4)^2}{29.2}=87.2-0.66=86.54$$

(vi) Between group adjusted sum of squares $(SSy\cdot x_b)$

$$=\text{Total adjusted } SS-\text{Within adjusted } SS$$

$$=118.25-86.54=31.71$$

Table 23.11: Summary of covariance analysis

Source of variation	SS	df	MS	F
Between groups	31.71	$k-1=3-1=2$	15.85	
Within groups (error)	86.54	$k(n-1)-1$ $=3(5-1)-1=11^{*}$	7.86	2.01
Total	118.25	13		

$$F_{y\cdot x}=\frac{15.85}{7.86}=2.01F\,(df\,2,11)\ \begin{matrix}3.98 & (0.05)\\ 7.20 & (0.01)\end{matrix}$$

Steps of computation

Step 1: ANOVA of *X* scores

In this step, ANOVA of *X* scores has been carried out separately. The outcome of this ANOVA is summarized in Table 23.8. The outcome, that is, F ratio shows that the three groups differ significantly. $F(df=2,12)=7.10$ is more than *F* value of 0.01 level, so it is significant at 0.01 level.

Step 2: ANOVA of Y scores

In this second step, ANOVA of *Y* scores has been worked out separately. The outcome of this ANOVA of *Y* scores has been presented in Table 23.9. It is clear that the three groups also differed

---

[*] 1 *df* is lost because of regression of Y on X.

significantly on the criterion measure Y as the obtained F, that is, 5.29 exceeds the value of F at 0.05 (df 2, 12). Again, it is concluded that the three groups differed on the variate of Y (criterion)

Step 3 : Analysis of covariance

Here ANCOVA has been worked out. First, the total sum of products (XY) has been calculated and this sum of products has been divided into two components—sum of products between groups and sum of products within groups. Subsequently, computations have been done for the purpose of correcting Y scores for differences in X scores. The symbol SS_{yx} has been used to refer to the adjusted scores for any variability in Y contributed by X scores or that variability in X is held constant. For convenience, the adjusted sum of squares of X, adjusted sum of squares of Y and adjusted sum of squares for products (XY) have been summarized in Table 23.10. It is observed that the obtained $F = 2.01$ is not significant at 0.05 level. Thus analysis of variance for adjusted Y scores does not show a significant difference between the groups. Hence, Ho (null hypothesis) is accepted with the conclusion that the groups don't differ significantly after giving treatments. Hence, out of the three training techniques, no one technique can be said to be better than another.

If we pay attention to the above numerical example, it is clear that the three groups differ significantly with respect to X as well as Y scores. However, when Y scores are adjusted to account for correlation between two variables, the groups no longer differ significantly. It appears therefore, that the former difference of Y scores was simply a reflection of differences in X scores, that is, there is regression Y on X. Let us find if regression of Y and X is really significant.

The total sum of squares of Y measures are divided into two components—one corresponding to those parts of Y scores which can be predicted from the X scores and the other (residual) corresponding to those parts of Y scores, which are independent of X.

$$\text{Sum of squares (SS) due to regression} = \frac{(SS_t)^2}{SS_{xt}}$$

$$= \frac{(54)^2}{63.73} = \frac{2916}{63.73} = 45.75$$

Total SS of $Y = 164$

$\therefore$ Residual $= 164 - 45.75 = 118.25$

Out of the total variation of Y scores of 164, 45.75 is due to those parts of Y scores which can be predicted from X scores and 118.25 corresponds to those parts of Y scores which are independent of X scores. The test of significance of the regression of Y on X is being presented in Table 23.12

Table 23.12. ANOVA testing the significance of regression of Y on X

Source of variation	SS	df	MS	F
Due to regression	45.75	1	45.75	5.03
Residual (error)	118.25	13	9.09	
Total	164	14		

$$F = \frac{45.75}{9.09} = 5.03 \quad (P < 0.05)$$

F(df 1, 13) = 4.65 (0.05 level)

F(df 1, 13) = 9.07 (0.01 level)

Since obtained F exceeds the critical value of 4.65 at 0.05 level for df 1 and 13, H_0 (null hypothesis) is rejected, we conclude that the regression of Y on X is significant. This shows that the X and Y scores are significantly associated.

In general, when the correlation between X and Y cores is high and among the means low, ANCOVA will often produce a significant F. In the above example the correlation between X and Y is not high ($r_w = 0.08$) and correlation among means is quite high ($r_b = 0.96$) and therefore, the outcome, that is, F for ANCOVA is not significant

$$r_w = \frac{SP_w}{\sqrt{SS_{xw} \cdot SS_{yw}}} = \frac{4.4}{\sqrt{29.2 \times 87.2}} = \frac{4.4}{50.4} = 0.08$$

$$r_b = \frac{SP_b}{\sqrt{SS_{xb} \cdot SS_{yb}}} = \frac{49.6}{\sqrt{34.53 \times 76.8}} = \frac{49.6}{51.49} = 0.96$$

So far as this example is concerned, the matter ends here. But suppose the obtained F of ANCOVA becomes significant, and therefore, H_0 is rejected. In such a case, our objective then will be to proceed ahead and analyze the differences in final scores after correction or adjustment and conclude which one of these treatments is better than the others. Hence, another numerical example follows.

Numerical Example II

In this second example, we are going to discuss a different experiment where the groups are not significantly different on X as well as Y variates but the adjusted F (F_{yx}) is significant. This example illustrates a situation which is opposite to the first numerical example. The data on this fictitious experiment are as under:

Table 23.13. Scores on a covariate (X) and a criterion variate (Y) for the single factor experiment

Group A		Group B		Group C	
Initial scores X_1	Final scores Y_1	Initial scores X_2	Final scores Y_2	Initial scores X_3	Final scores Y_3
5	10	15	15	10	15
10	20	5	10	25	25
15	25	10	15	10	20
20	30	25	30	5	10
10	20	15	20	5	10
Sum: 60	105	70	90	55	80
Mean: 12	21	14	18	11	16

$$\Sigma X = 60 + 70 + 55 = 185$$

$$\Sigma X = 105 + 90 + 80 = 275$$

The computational work for ANCOVA for the data presented in Table 23.13 may be started by arranging the above data as in Table 23.14.

Table 23.14 Computational Arrangement for ANCOVA

Group A				
X_1	Y_1	$(X_1)(Y_1)$	X_1^2	Y_1^2
5	10	50	25	100
10	20	200	100	400
15	25	375	225	625
20	30	600	400	900
10	20	200	100	400
Sum = 60	105	1425	850	2425
Means = 12	21			

Group B				
X_2	Y_2	$(X_2)(Y_2)$	X_2^2	Y_2^2
15	15	225	225	225
5	10	50	25	100
10	15	150	100	225
25	30	750	625	900
15	20	300	225	400
70	90	1475	1200	1850
14	18			

	Group C				
	X_3	Y_3	$(X_3)\ (Y_3)$	X_3^2	Y_3^2
	10	15	150	100	225
	25	25	625	625	625
	10	20	200	100	400
	5	10	50	25	100
	5	10	50	25	100
	Sum: 55	80	1075	875	1450
	Mean: 11	16			

$\Sigma X = 60 + 70 + 55 = 185$

$\Sigma Y = 105 + 90 + 80 = 275$

Steps of computation

A. ANOVA of *X* scores:

(i) Correction $(C_x) = \dfrac{(\Sigma X)^2}{N} = \dfrac{(185)^2}{15} = 2281.67$

(ii) Total sum of squares $(SS_{tx}) =$

$$\Sigma X^2 - C_x = (850 + 1200 + 875) - 2281.67$$

$$= 2925 - 2281.67 = 643.33$$

(iii) Between group sum of squares $(SS_{xb}) =$

$$\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} - C$$

$$= \frac{(60)^2}{5} + \frac{(70)^2}{5} + \frac{(55)^2}{5} - 2281.67 = 2305 - 2281.67 = 23.33$$

(iv) Within group sum of squares $(SS_{xw}) = SS_{tx} - SS_{xb}$

$$= 643.33 - 23.33 = 620$$

Table 23.15. Summary of ANOVA of *X* scores

Source of variation	*SS*	*df*	*MS*	*F*
Between groups	23.33	$k-1=3-1=2$	11.66	
Within groups (error)	620	$N-k=15-3=12$	51.66	0.22
Total	643.33	14		

$F_x = \dfrac{11.66}{51.66} = 0.22$ $\quad$ F at 0.05 3.88 $\quad$ F at 0.01 6.93 $\quad$ $F(df2, 12)$

B. ANOVA of *Y* scores

(i) Correction $(C_y) = \dfrac{(\Sigma Y)^2}{N} = \dfrac{(105+90+80)^2}{15} = \dfrac{(275)^2}{15} = 5041.66$

(ii) Total sum of squares $(SS_{ty}) = \Sigma Y^2 - C_y$

$$= (2425 + 1850 + 1450) - 5041.66$$

$$= 5725 - 5041.66 = 683.34$$

(iii) Between group sum of squares (SS_{yb})

$$= \frac{(\Sigma Y_1)^2}{n_1} + \frac{(\Sigma Y_2)^2}{n_2} + \frac{(\Sigma Y_3)^2}{n_3} - C_y$$

$$= \frac{(105)^2}{5} + \frac{(90)^2}{5} + \frac{(80)^2}{5} - 5041.66$$

$$= 5105 - 5041.66 = 63.34$$

(iv) Within group sum of squares $(SS_{yw}) = SS_{ty} - SS_{yb}$

$$= 683.34 - 63.33 = 620.00$$

Table 23.16. Summary of ANOVA of Y Scores

Source	*SS*	*df*	*MS*	*F*
Between groups	23.33	$k-1=3-1=2$	31.67	
Within groups (error)	620	$N-k=15-3=12$	51.67	0.61
Total	643.33	14		

$F = \dfrac{31.67}{51.67} = 0.61$ $\quad$ F at 0.05 3.88 $\quad$ F at 0.01 6.93 $\quad$ $(df2, 12)$

C. Analysis of covariance

(i) Total sum of products (SP_t)

$$= (\Sigma X_1Y_1 + \Sigma X_2Y_2 + \Sigma X_3Y_3) - \frac{\Sigma X \cdot \Sigma Y}{N}$$

$$= (1425 + 1475 + 1075) - \frac{(185 \times 275)}{15}$$

$$= 3975 - 3391.66 = 583.34$$

(ii) Sum of products between groups (SP_b)

$$= \frac{(\Sigma X_1 \cdot \Sigma Y_1)}{n_1} + \frac{(\Sigma X_2 \cdot \Sigma Y_2)}{n_2} + \frac{(\Sigma X_3 \cdot \Sigma Y_3)}{n_3} - \frac{\Sigma X \cdot \Sigma Y}{N}$$

$$= \frac{(60 \times 105)}{5} + \frac{(70 \times 90)}{5} + \frac{(55 \times 80)}{5} - \frac{(185 \times 275)}{15}$$

$$= 3400 - 3391.66 = 8.34$$

(iii) Sum of Products within groups $(SP_w) = SPt - SP_b$

$$= 583.34 - 8.34 = 575$$

For convenience, the total sum of products (SP_t) may be arranged with sum of squares for both X and Y variables as under:

Table 23.17. Summary of sum of squares of *X* and *Y* and that of sum of products

	Total	**Within**	**Between**
Sum of products (XY)	583.34	575	8.34
Sum of squares (X)	643.33	620	23.33
Sum of squares (Y)	683.40	620.07	63.33

(iv) Adjusted SS for $Y (SS_{y.x})$

$$\text{Total adjusted } SS = SS_{yt} - \frac{(SP_t)^2}{SS_{xt}}$$

$$= 683.40 - \frac{(583.34)^2}{643.33}$$

$$= 683.40 - 528.94 = 154.46$$

$$\text{Within adjusted } SS = SS_{yw} - \frac{(SP_w)^2}{SS_{xw}}$$

$$= 620.07 - \frac{(575)^2}{620} = 620.07 - 533.26 = 86.81$$

Between adjusted SS = Total – within

$$= 154.46 - 86.81 = 67.65$$

Table 23.18. Summary of covariance analysis

Source of variation	*SS*	*df*	*MS*	*F*
Between groups	67.65	$k-1=3-1=2$	33.82	
Within groups (error)	86.81	$k(n-1)-1$ $=3(5-1)-1=11^{*}$	7.89	4.29
Total	154.46	13		

$$F_{y.x} = \frac{33.51}{7.94} = 4.29 \qquad F(df\ 2, 13)\ 0.05 = 3.80$$

Discussion

Steps involved in the computation of Numerical Example II are essentially the same as those involved in the computation of Numerical Example I. Therefore, those steps are not being explained here.

It is further observed that F test applied to the initial scores (X) is not significant even at 0.05 level. It means that the three groups don't differ significantly on the X test. Likewise, F test applied to the final scores (Y) is also not significant at 0.05 level. Again, it is concluded that the three groups don't differ significantly on Y test also. However, the analysis carried out on Y measure by adjusting the variability contributed by the initial X scores indicated a significant result. In other words, the adjusted F of 4.24 is found to be significant at 0.05 level. Thus the three groups differ significantly when the groups have been adjusted for initial differences in X. Hence, the comparisons among means of the groups must be tested.

But before testing mean differences of the three groups, let us determine whether the regression of Y on X is significant. For this, the total sum of squares of Y measures are divided into two components—one corresponding to those parts of Y scores which can be predicted from the X scores and the other (residual) corresponding to those parts of Y scores which are independent of X. The sum of squares that can be predicted from the X scores is obtained by the formula: $\frac{SP_t^{\ 2}}{SS_{tx}}$.

$$\text{Sum of squares due to regression} = \frac{(583.34)^2}{643.33} = 528.94$$

$$\text{Total sum of squares of } Y = 683.40$$

$$\text{Residual} = 683.40 - 528.94 = 154.46$$

Table 23.19 ANOVA for the regression of *Y* on *X*

Source	*SS*	*df*	*MS*	*F*
Due to regression	528.94	1	528.94	44.52
Residual	154.46	13	11.88	
Total	683.40	14		

$$F = \frac{528.94}{11.88} = 44.52 \qquad \begin{matrix} F \text{ at } 0.05 & 4.67 \\ F \text{ at } 0.01 & 9.07 \end{matrix} \ (df = 2, 13)$$

Out of the total variation of Y scores of 683.40, 528.94 is due to those parts of Y scores, which can be predicted from X scores and 154.46 is due to those parts of Y scores which is, in fact, independent of X. The test of significance has been carried out in Table 23.19. The obtained F value is 44.52, which far exceeds the critical value of 9.07 for df 1 and 13 at 0.01 level. Thus H_0 (null hypothesis) is rejected and it is concluded that the regression of Y on X is highly significant. This provides an evidence for the fact that X and Y scores are significantly associated.

Now let us test the mean differences of the three groups by using t test:

The individual Y means can be adjusted for differences in X means by the following formula:

$$M_{y.x} = M_y - bw\,(M_x - G_{Mx}) \tag{23.24}$$

where,

$M_{y.x}$ = Adjusted mean of Y when X is kept constant

M_y = Mean of uncorrected Y scores

M_x = Mean of X scores

bw = Regression coefficient of within groups (SP_w / SS_{xw})

G_M = Grand mean of X scores

$$\text{Here, } bw = \frac{SP_w}{SS_{xw}} = \frac{575}{620} = 0.93$$

---

[*] 1 df is lost because of regresson of Y on X.

$$G_{Mx} = \frac{G_x}{N} = \frac{185}{15} = 12.33$$

Now by applying formula 23.24, the adjusted Y means for each of three groups can be calculated as given below.

For Group A:

$$M_{y.x} = 21 - 0.93\,(12 - 12.33)$$
$$= 21 - 0.93\,(-33) = 21 + 0.306 = 21.306 = 21.31\,(M_1)$$

For Group B:

$$M_{y.x} = 18 - 0.93(14 - 12.33)$$
$$= 18 - 1.55 = 16.45\,(M_2)$$

For Group C:

$$M_{y.x} = 16 - 0.93\,(11 - 12.33)$$
$$= 16 - 0.93(-1.33) = 16 + 1.24 = 17.24\,(M_3)$$

Now we proceed to test differences between the three means by computing t test. The formula is:

$$t = \frac{\text{mean difference}}{SE_D}$$

SE_D of any two adjusted means can be computed by using the following formula:

$$SE_D = \sqrt{MS_w\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \qquad (23.25)$$

MS_w = Mean squares of within groups (adjusted)

N_1 = Number of subjects in the first group compared

N_2 = Number of subjects in the second group compared

Here, $df = k(n-1) - 1 = 3\,(5-1) - 1 = 12 - 1 = 11$

where, k = number of groups to be compared

n = number of subjects in each group being compared

$$SE_D = \sqrt{7.89\left(\frac{1}{5} + \frac{1}{5}\right)} = \sqrt{3.156} = 1.77$$

The computed value of t is:

(i) $\frac{M_1 - M_2}{SE_D} = \frac{21.31 - 16.45}{1.77} = \frac{4.86}{1.77} = 2.74$ (Significant at 0.05 level) $df = 11$

(ii) $\frac{M_2 - M_3}{SE_D} = \frac{16.45 - 17.24}{1.77} = \frac{0.79}{1.77} = 0.45$ (Not significant) $df = 11$

(iii) $\frac{M_1 - M_3}{SE_D} = \frac{21.31 - 17.24}{1.77} = \frac{4.07}{1.77} = 2.30$ (Significant at 0.05 level) $df = 11$

The obtained results show that Group A and Group B differ significantly. Likewise, Group A and Group C differ significantly. However, Group B does not differ significantly from Group C.

In Table 23.20, an example of computation of F test from repeated measures (within-group design) has been provided. A researcher has examined the sales performance of 5 new employees in a building construction company. To see if there is significant trend towards improvement, the number of homes sold is recorded each month for 3 months of employment. The researcher wants to answer the question: Is there a significant change in sale performance with more work experience?

Table 23.20. Computation of repeated measures ANOVA

Person	Month 1	Month 2	Month 3	Total
A	1	3	4	8
B	4	6	8	18
C	3	3	5	11
D	2	4	7	13
E	0	4	6	10
	Sum: 10	20	30	60
	$\overline{X}$ (or mean): =2	4	6	

Computation:

(i) Correction Term $(C) = \frac{(60)^2}{15} = 240$

(ii) Total SS(Sum of squares) $= (1^2 + 4^2 + 3^2 + \ldots + 5^2 + 7^2 + 6^2) - C$

$$= 306 - 240 = 66$$

(iii) Between subject $SS = \left(\frac{8^2}{3} + \frac{18^2}{3} + \frac{11^2}{3} + \frac{13^2}{3} + \frac{10^2}{3}\right) - C$

$$= 259.3 - 240 = 19.3$$

(iv) Within subject SS = Total SS – Between subject SS

$$= 66 - 19.3 = 46.7$$

(v) Between treatment $SS = \left(\frac{10^2}{5} + \frac{20^2}{5} + \frac{30^2}{5}\right) - C$

$$= 280 - 240 = 40$$

(vi) Residual (Error) = Within subject SS – Between treatment SS

$$= 46.7 - 40 = 6.7$$

Table 23.21. Summary of ANOVA for single factor exepriment with repeated measures

Source of variation	SS	df	MS	F
Between SS	**19.3**	$k-1=3-1=2$		
Within SS	**46.7**	$N-k=15-3=12$		
Treatment SS	40	$n-1=5-1=4$	10	11.90
Residual (error)	6.7	$(N-k)-(n-1)$ $=12-4=8$	0.84	
Total	66	14		

$$F = \frac{\text{Between treatment } MS}{\text{Error } MS} = \frac{10}{.84} = 11.90$$

df for numerator = 4 $\qquad$ F at 0.05 = 3.84

df for denominator = 8 $\qquad$ F at 0.01 = 7.01

In summary of ANOVA the results of the computation have been summarized in Table 23.21. The between subject SS and within subjects SS, were 19.3 and 46.7 and $k-1=3-1=2$ and $N-k=15-3=12$ (under df) are all set in bold type. This indicates that the total sum of squares has only two components—between subjects and within subjects sum of squares. The within subjects sum of squares is *further* divided into treatment sum of squares and residual sum of squares. The F ratio is obtained by dividing between treatment MS by MS residual (error). The obtained F is significant and therefore, H_0 (null hypothesis) is rejected.

Post Hoc test with Repeated measures ANOVA

To determine exactly where significant difference exists, the researcher must follow the ANOVA with post hoc tests. Like independent measures ANOVA, in repeated measures ANOVA Tukey's HSD can be computed as post hoc test. The formula for Tukey's HSD in case of repeated measures ANOVA becomes:

$$HSD = q\sqrt{\frac{MS_{\text{error}}}{n}} \qquad (23.26)$$

where q is the value of the Studentized Range statistic, MS_{error} is residual variance and n is the number of scores in each treatment.

In this experiment MS_{error} is 0.84 with $df = 8$. At 0.05 level of significance with the $k = 3$, the value of q for these data with the help of Table 23.5 is 4.04.

$$HSD = 4.04\sqrt{\frac{0.84}{5}} = 1.66$$

It means that the mean difference between any two samples must be at least 1.66 to be significant. Using this value, we can make the following conclusions:

Mean difference between 2 months and 1 month = 4 – 2 = 2 (significant)

Mean difference between 3 months and 1 month = 6 – 2 = 4 (significant)

Mean difference between 3 months and 2 months = 6 – 4 = 2 (significant)

All these three mean differences are significant in light of Tukey's HSD test that determines the minimum difference between two treatment means that is necessary for being significant.

4. Pearson *r*

Of all the measures of correlation the Pearson *r*, named after Prof. Karl Pearson, is one of the most common methods of assessing the association between two variables under study. The Pearson correlation measures the degree and direction of linear relationship between two interval-level variables. Pearson *r* represents the extent to which the same individuals, events, etc., occupy the same relative position on two variables. It is a symmetric statistic that makes no distinction between independent variable and dependent variable. It is also known as Pearson product–moment correlation and abbreviated to *r*. The size of Pearson *r* varies from +1 through 0 to –1. In fact, all correlation coefficients have the limit of +1 and –1. A coefficient of +1 indicates perfect positive correlation, and a coefficient of –1 indicates perfect negative correlation. The coefficient of correlation tells us two things. First, it indicates the magnitude of relationship. A correlation coefficient of, say, +0.90 or –0.90 gives the same information about the magnitude or size of correlation. *The sign makes no variation in the size of the correlation.* Second, it gives indication regarding the direction of the correlation coefficient. A positive correlation indicates a

---

similar trend of relationship between two variables, that is, as one increases, the other also increases or as one decreases, the other also decreases. Consider the relationship between intelligence test scores and classroom achievement. Generally, as intelligence test scores are raised, classroom achievement is also raised. And, therefore, the direction of the correlation between these two variables is positive. Likewise, consider the correlation between fatigue and output. As fatigue increases, output decreases. Here the relationship is negative because as one increases, the other decreases. Sometimes, the relationship is not consistent. And in this situation coefficient of correlation is likely to be zero.

The Pearson product–moment correlation has two important assumptions.

1. The relationship between *X* and *Y* variables should be linear. A linear relationship refers to the tendency of the data, when plotted, to follow a straight line as closely as possible. Although there are some statistical tests through which one can test whether or not the relationship is linear, generally this is determined by the inspection of the scatter diagram or correlation table.

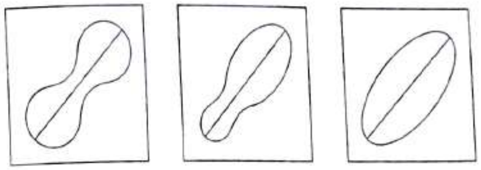
(a) (b) (c)

Fig. 23.7 (a) and (b) nonhomoscedastic (c) homoscedastic and linear

2. The second assumption is of *homoscedasticity* (*homo* means 'like' and *scedasticity* means scatteredness). Defined statistically, we can say that homoscedasticity refers to the fact that standard deviations (or variances) for columns and rows in the scatter diagram are equal or nearly equal. Figure 23.7 illustrates homoscedastic and non-homoscedastic distributions.

In this figure we have three diagrams. In diagram (a) the variance of the distribution near the centre is smaller than the variances near both extremes, and hence, the distribution is non-homoscedastic. In diagram (b) the variances near the bottom extreme are lower than the variances at middle or at top extreme, and, therefore, the distribution is non-homoscedastic. In diagram (c), the variances are equal throughout as well as linear.

As we know, the Pearson correlation assesses the degree and direction of linear relation between two interval-level variables. Pearson correlation identified by letter *r* is computed by:

$$r = \frac{\text{degree to which } X \text{ and } Y \text{ vary together}}{\text{degree to which } X \text{ and } Y \text{ vary separately}} \qquad (23.27)$$

$$\text{or} \qquad r = \frac{\text{covariability of } X \text{ and } Y}{\text{variability of } X \text{ and } Y \text{ separately}}$$

In any distribution when there is a perfect linear relationship, every change in *X* is accompanied by a corresponding change in *Y*. In such a case, the covariability perfectly reflects the total variability of *X* and *Y* separately. The overall result is a perfect correlation of 1.00. On the other hand, when there is no linear relationship, a change in *X* variable does not produce any predictable change in *Y*. In this case, there will be no covariability and the resulting correlation is zero. Pearson *r* can be calculated by several formulas. In this book, we shall illustrate the calculation of Pearson r by the *raw score formula* or *machine formula*. The equation is:

$$r = \frac{N\Sigma XY - \Sigma X\,\Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \qquad (23.28)$$

where, r = Pearson product–moment correlation coefficient; N = number of scores; X = scores in X variable; and Y = scores in Y variable.

Table 23.22 presents the scores of 10 students who were administered intelligence test (X) and anxiety test (Y). Pearson r has also been calculated from the ten pairs of scores. The significance of the obtained r is tested with the help of a table (see table given in Appendix F, Downie & Heath 1970, 378). The obtained value of r is less than the value required at even 0.05 level of significance. Hence, the null hypothesis is accepted and it is concluded that the scores on intelligence test and the scores on the anxiety test are not correlated. The sign makes no difference in the magnitude of the correlation.

When the data are arranged in bivariate distribution as is the case in the scatter diagram or in the correlation table, Pearson r should be calculated by the following formula:

$$r = \frac{\frac{\Sigma x'y'}{N} - C_x C_y}{(\sigma_{x'})(\sigma_{y'})} \tag{23.29}$$

where, r = Pearson r; x' = deviation of scores from mean on X test; y' = deviation of scores from mean on Y test; N = sum of frequencies; C_x = correction in X-series scores; and C_y = correction in Y-series scores.

Table 23.22 Pearson *r* by raw-score method

X	Y	X^2	Y^2	XY
10	8	100	64	80
5	20	25	400	100
6	15	36	225	90
3	13	9	169	39
8	16	64	256	128
12	20	144	400	240
13	13	169	169	169
20	11	400	121	220
15	10	225	100	150
10	12	100	144	120
$\Sigma X = 102$	$\Sigma Y = 138$	$\Sigma X^2 = 1272$	$\Sigma Y^2 = 2048$	$\Sigma XY = 1336$

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2 2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{(10)(1336) - (102)(138)}{\sqrt{[(10)(1272) - (102)^2][(10)(2048) - (138)^2]}}$$

$$= \frac{-716}{\sqrt{3325776}} = \frac{-716}{1823.671} = -0392 = -039$$

$$df = N - 2 = 10 - 2 = 8$$

While interpreting correlation coefficient, the researcher must take into account two different circumstances, which can cause a higher or lower correlation. One circumstance arises when one person or relatively few persons have pair of scores that are markedly different from the rest of the sample's scores. Such individual's pair of scores are called as *outliers*. In fact, the actual magnitude of correlation is strongly affected by one or more outliers. The second circumstance is one where all other things are equal, and there is a homogeneity in group of scores. In such a situation, the magnitude of Pearson r will be lower one. In simple words, if there is smaller range of scores, there will be smaller value of r.

Researchers have pointed out many ways for interpreting a correlation coefficient, depending upon the purpose of the researchers and the circumstances that influence the correlation's magnitude. One popular method uses a *crude criterion* for interpreting the magnitude of a correlation. This is presented below.

Value of *r*	**Interpretation**
0.80 to 1.00	High to very high
0.60 to 0.80	Substantial
0.40 to 0.60	Moderate
0.20 to 0.40	Low
0.00 to 0.20	Negligible

Another method of interpreting correlation coefficient is on the basis of *statistical significance* of correlation based on the concept of sampling error. Still another way of interpreting correlation coefficient is in terms of variance. In fact, the variance of the scores that the researcher wants to predict is divided into two parts—one that is explained by predictor or treatment variable and the other that is explained by other factor (generally unknown) including sampling error. The researcher finds the percentage of explained variance by calculating r^2 popularly called as *coefficient of determination*. Then, percentage of variance not explained by the predictor variable is $1 - r^2$. Let us take an example to illustrate this fact. Suppose the researcher wants to predict general academic achievement (Y variable) on the basis of IQ (X variable) and he obtains r of 0.65 between IQ and general academic achievement. He can use this correlation to find $r^2 = (0.65)^2 = 0.42$. This means that 42 per cent of variance in general academic achievement is predictable from the variance of IQ. It also means that 58 per cent of the variance of general academic achievement is due to factors other than IQ such as home environment, schools environment, level of motivation, etc.

Still another way of interpreting Pearson r is in terms of standard error of estimate and *coefficient of alienation* denoted by letter K. As we know, in predicting X on the basis Y or Y on the basis of X, we also compute standard error of estimate for gauging the predictive strength. In fact, standard error of estimate is the standard deviation of the difference between the actual values of the predicted variable and those estimated from regression equation. Briefly, it is simply the standard deviation of the errors of estimate. The formula for standard error of estimate is:

$$SE(est)_{y.x} = \sigma_y \sqrt{1 - r_{xy}^2} \quad (Y \text{ on the basis of } X) \tag{23.30}$$

$$SE(est)_{x.y} = \sigma_x \sqrt{1 - r_{xy}^2} \quad (X \text{ on the basis of } Y) \tag{23.31}$$

Researchers have shown that it is possible to gauge from $\sqrt{1 - r^2}$ the predictive strength of an r. This $\sqrt{1 - r^2}$ is called as *coefficient of alienation* and is symbolized by letter K.

$$K = \sqrt{1 - r^2} \tag{23.32}$$

K measures *absence* of relationship between two variables just as r measures presence of relationship between two variables. In simple words when $K = 1$, $r = 0.00$ and when $K = .00$ $r = 1.00$. This obviously means that the larger the value of *K*, the smaller the extent of relationship and less precise and accurate will be the forecast from *X* to *Y* and the vice versa.

In a nutshell, it can be said that coefficient of alienation provides us a measure of the extent to which errors in our estimates (while making prediction) are reduced. When $r = \pm 1.00$, the coefficient of alienation becomes 0.00 and, in fact, we make no errors in predicting one score from the other. When $r = 0.00$, the coefficient of alienation is 1.00 and therefore, the standard error of estimate becomes the same as standard deviation of marginal frequencies. When $r = 0.50$, the *K* (coefficient of alienation) is 0.87, which indicates that our error of estimate are 0.87 as large as or 13% (100% – 87% = 13%) smaller than they would be if there were no correlation between *X* and *Y* variables.

There are three important *applications* of correlation.

(i) *Prediction:* If two variables are correlated in some systematic way, it is possible to use one of the variables for making accurate prediction about the other. If the correlation is perfect (±1.00), this prediction is hundred per cent accurate.

(ii) *Validity of the test:* Frequently, the researcher develops a new test and he wants to validate it. In other words, he wants to know if the test is truly measuring what it claims to measure. One common way of knowing about this validity is to correlate the new test with some other test that is measuring what the new test is measuring. For example, a newly developed intelligence test may be correlated against another intelligence test developed earlier. If the correlation is high, it is said that the newly developed intelligence test has sufficient degree of validity.

(iii) *Theory verification:* Correlation is also used in the task of theory verification. Many psychological theories make specific predictions about the relationship between two variables. For example, social psychologist may develop a theory predicting personality type and prosocial behaviour. Likewise, physiological psychologist may predict a relationship between brain size and learning ability. In each case, the prediction of the theory could be tested by calculating correlation between the two concerned variables.

5. Partial correlation and Multiple correlation

When *X* and *Y* become correlated with each other, there is always the possibility that this correlation is due to the association between each of the two variables and a third variable. For example, among a group of schoolchildren of different ages, the researcher may find a high correlation between size of vocabulary and height. In fact, this correlation may not display a genuine correlation between size of vocabulary (*X*) and height (*Y*) but may also result from the fact that both vocabulary size and height may be related with a third variable that is, age (*Z*). Likewise, if the researcher has found a high correlation between academic achievement and the time devoted to study, it may not be true because this may result from the fact that both variables may be related to a third variable, that is, intelligence.

In designing any experiment, the researcher has the alternative of either introducing *experimental control* for eliminating the influence of third variable or using *statistical methods* to control or eliminate the influence of the third variable. Suppose the experimenter wishes to study relation between ability to memorize (*X*) and ability to solve certain kinds of problems (*Y*). Both these variables are related to intelligence (*Z*). Therefore, for determining genuine correlation between ability to memorize and ability to solve problems, it is essential that the third factor, that is, intelligence must be controlled. If the experimenter wants to control intelligence through experimental control, he might choose subjects with equal intelligence. But somehow, if the experimental control is not feasible then statistical control can be applied.

Partial correlation is one such statistical method that allows the researcher to have a correlation between two variables by controlling the impact of the third variable.

In simple words, partial correlation coefficient is a measure of *linear relationship* between two interval-level variables, controlling for one or more other variables. Controlling a variable is also called *partialing out* or *adjusting* for or *holding* a variable constant. The meaning is the same. Like an ordinary correlation coefficient, partial coefficient ranges from + 1.00 to –1.00. When the correlation between *X* and *Y* is found by controlling only one other variable, it is called as *first-order partial correlation* but when two variables are controlled, it is called as *second-order partial correlation*.

Let us give a brief look on the logic of partial correlation coefficient. Technically, partial correlation between variables *X* and *Y* controlling for variable *Z* is the correlation of the residuals of the regressions of *X* on *Z* and *Y* on *Z*. In simple words, one can think of control variable (*Z*) serving as independent variable for predicting *X* and *Y* in two separate regression analyses. In this process, two sets of residuals* are computed. One set of residuals represents the variation in *X* not explained by *Z* and the other set of residuals represent the variation in *Y* not explained by *Z*. The partial correlation coefficient between *X* and *Y* controlling for *Z* can be easily computed by correlating these two sets of residuals.

The formulas for computing partials correlation are as under.

First-order partial correlation:

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{(1-r_{xz}^2)(1-r_{yz}^2)} \tag{23.33}$$

Second-order partial correlation:

$$r_{xy \cdot zA} = \frac{r_{xy \cdot z} - r_{XA \cdot Z} \cdot r_{YA \cdot Z}}{(1-r_{XA \cdot Z}^2)(1-r_{YAZ}^2)} \tag{23.34}$$

(Computational equations for higher-order partial correlation coefficients will not be discussed here. The order of partial correlation coefficient indicates the number of variables controlled.)

where, X = Independent variable;
Y = Dependent variable;
Z = Control variable;
A = Control variable.

Let us take an example to illustrate the computation of first–order partial correlation.

(a) Zero-order correlation between ability to memorize (*X*) and ability to solve certain problems $(Y) = r_{xy} = 0.80$

(b) Zero-order correlation between *X* and intelligence $(Z) = r_{xz} = 0.05$

(c) Zero-order correlation between *Y* and $Z = r_{yz} = 0.60$

$$r_{xy \cdot z} = \frac{0.80 - (0.50)(0.60)}{\sqrt{(1-0.50^2)(1-0.60^2)}}$$

$$= \frac{0.80 - 0.30}{\sqrt{(0.75)(0.64)}} = \frac{0.50}{0.69} = 0.72$$

---

* Residuals mean the size and direction of error in prediction.

This shows that if all subjects would have been of *equal* intelligence, then correlation between ability to memorize and ability to solve certain problems would be 0.72 rather than 0.80.

However, sometimes situation may arise where control variable (*Z* variable) is completely removed from one variable only, say from *Y*, and then, *X* may be correlated with the residual of Y. This type of correlation is called as *part correlation* or *semipartial correlation*. The part correlation between X and the residuals of *Y* and with *Z* removed, is given by:

$$r_{x(y.z)} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{yz}^2}} \tag{23.35}$$

If $r_{xy} = 0.75, r_{xz} = 0.80$ and $r_{yz} = 0.75$, then the part correlation is:

$$r_{x(y.z)} = \frac{0.75 - (0.80)(0.75)}{\sqrt{1 - 0.72}} = \frac{0.15}{0.53} = 0.283$$

Application of part correlation generally arises where measurements (or scores) have been obtained prior to giving the experimental treatment and subsequent measurements on the same subjects or participants under experimental treatment.

Multiple correlation

Multiple correlation coefficient denoted by letter *R*, in its simplest form, is understood as the relationship between one variable and a combination of at least two other variables. The multiple correlation coefficient (*R*) is defined as a measure of linear relationship between a dependent variable and the combined effects of two or more independent variables. In simple words, *R* estimates the correlation between scores actually earned on the criterion variable and scores predicted on the criterion variable from two or more predictor variables in multiple regression equation. Making predictions in this situation is called as *multiple regression*.

When we want to know the *R* between *Y* (dependent variable) and the combined effects of X_2 X_3 (two independent variables), the formula for multiple correlation coefficient becomes like this:

$$R_{1.23}\sqrt{\frac{r_{12}^2 + r_{13}^2\,(2r_{12}r_{13}r_{23})}{1 - r_{23}^2}} \tag{23.36}$$

Let us take an example.

Variable 1: Academic achievement

Variable 2: Study habit

Variable 3: Intelligence

Zero order correlation between these three variables are as under:

$$r_{12} = 0.50$$

$$r_{13} = 0.60$$

$$r_{23} = 0.40$$

$$R = \sqrt{\frac{(0.50)^2(0.60)^2 - 2\,(0.50)\,(0.60)\,(0.40)}{1 - (0.40)^2}}$$

$$= \sqrt{0.4405} = 0.66$$

Multiple correlation coefficient *R* is interpreted in the same way as correlation coefficient *r*. It is a measure of relation between dependent variable and the combined effect of the set of independent variables. Just as r^2 is a proportion, so also R^2 is a proportion. When *R* is squared, that is, R^2, it is called *coefficient of multiple determination*, which indicates the proportion of total variance in the dependent variable that is explained jointly by two or more independent variables. In the above example, *R* is .66 and therefore $R^2 = 0.4356$ or 0.44. It means that 44% of variation in academic achievement has been explained by study habit and intelligence and remaining 56% has not been explained. Note that R^2 is *never* smaller than any single r^2 between dependent variable and independent variable.

Correlation and Causation

A very common error in interpreting correlation is to *assume* that a correlation necessarily implies a cause-and-effect relationship between two variables. Our lives are constantly bombarded with reports of relationship between two variables. A few examples are: cigarette smoking is related to lungs disease; alcohol consumption is related to birth defects; hard labour during examination by students is related to good grade; carrot consumption is related to good eyesight. Do these relationships indicate cause-and-effect relationship between the two concerned variables? Does cigarette smoking causes lungs disease or carrot causes good eyesight? The answer is *no*.

What, in fact, is being said is that correltional studies simply do not allow the inference of causation. Correlation is a *necessary* but not a *sufficient* condition to establish a causal relationship between two variables. Faulty causal inferences from correlation data is called as *past hoc fallacy*.

NONPARAMETRIC STATISTICS

The important nonparametric statistics which have been included in this book are as follows:

1. Chi-square (X^2) test
2. Mann–Whitney *U* test
3. Rank-difference methods (both *rho* and *tau*)
4. Kendall's partial rank order correlation
5. Coefficient of concordance (*W*)
6. Median test
7. Kruskal–Wallis *H* test
8. Friedman test

Each of these techniques has been discussed below.

Chi-square (X^2) Test

The chi-square is one of the most important nonparametric statistics, which is used for several purposes. This test was originally developed by Karl Pearson and therefore also sometimes called as Pearson chi-square. Due to its smooth uses for various purposes, Guilford (1956) has called it the *general-purpose statistic*. It is a nonparametric statistic because it involves no assumption regarding the normalcy of distribution or homogeneity of the variances. The chi-square test is used when the data are expressed in terms of frequencies or proportions or percentages. The basic ideas behind use of X^2 are that (a) we have a theory of any kind concerning how our cases should be distributed, (b) we have a sample which shows how the cases actually are distributed, and (c) we want to know whether or not the differences between theoretical frequencies and observed frequencies are of such size that these differences might reasonably be attributed to the chance factor. The chi-square applies only to discrete data. However, any continuous data can be reduced to the categories in such a way that they can be treated as discrete data and then, the application of chi-square is justified. The formula for calculating X^2 is given in the next page.

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e} \qquad (23.37)$$

or,

$$X^2 = \Sigma \left(\frac{f_o^2}{f_e} \right) - N \qquad (23.38)$$

where, X^2 = chi-square; f_o = obtained or observed frequency; and f_e = expected frequency or theoretical frequency; N = total sum of observations.

The calculation of X^2 with formula 23.38 requires less arithmatic than calculation of X^2 with formula 23.37. Hence, it saves time.

There are several uses of the chi-square test.

First, chi-square may be used as a test of *equal probability hypothesis*. By equal probability hypothesis, we mean the probability of having the frequencies in all the given categories as equal. Suppose, for example,100 students answer an item in an attitude scale. The item has five categories of response options—*strongly agree, agree, neutral, disagree* and *strongly disagree*. According to the equal probability hypothesis, the expected frequency of responses given by 100 students would be 20 in each. The chi-square test would test whether or not the equal probability becomes tenable. If the value of the chi-square test is significant, the equal probability hypothesis becomes untenable and if the value of the chi-square is not significant, the equal probability hypothesis becomes tenable.

The second use of the chi–square test is in testing the significance of the *independence hypothesis*. By independence hypothesis is meant that one variable is not affected by, or related to, another variable and hence, these two varibales are independent. The chi-square is not a measure of the degree of relationship in such a situation. It merely provides an estimate of some factors other than chance (or sampling error), which account for the apparent relationship. Generally, in dealing with data related to *independence hypothesis*, they are first arranged in a contingency table. When observations on two variables are classified in a two-way table, data are called the *contingency data* and the table is known as the *contingency table*. Independence in a contingency table exists only when each tally exhibits a different event or individual.

The third important use of chi-square is in testing a hypothesis regarding the normal shape of a frequency distribution. When chi-square is used in this connection, it is commonly referred to as a test of goodness-of-fit.

The fourth use of chi-square is in testing the significance of several statistics. For example, for testing the significance of the phi-coefficient, coefficient of concordance and coefficient of contingency, we convert the respective values into chi-square values. If the chi-square value appears to be a significant one, we also take their original values as significant.

As an illustration, let us take an example of a 3 × 3 contingency table, which shows data of 200 students who were classified into three classes on the basis of their educational qualification (see Table 23.23). Their educational attainment is measured in course of study by classifying them as superior, average and inferior. Now, the question is: Is educational achievement related to educational qualification? The obtained data have been shown in Table 23.23. For the moment, omit the data given in parentheses because they indicate expected frequency and not observed frequency. The first step in calculating X^2 as a test of significance of independence or the relationship between educational achievement and educational qualification is to compute the expected frequency. The null hypothesis is that these two variables are not related or are independent, and if this hypothesis is true, the expected frequencies should be as follows.

Table 23.23 The use of Chi-square in a 3 × 3 contingency table

	Superior	Average	Inferior	
Master	30 (25)	15 (15)	5 (10)	50
Bachelor	25 (25)	10 (15)	15 (10)	50
Intermediate	45 (50)	35 (30)	20 (20)	100
	100	60	40	200

Cells of table	Expected frequency
Upper left	(100 × 50)/200 = 25
Upper middle	(60 × 50)/200 = 15
Upper right	(40 × 50)/200 = 10
Middle left	(100 × 50)/200 = 25
Middle middle	(60 × 50)/200 = 15
Middle right	(40 × 50)/200 = 10
Lower left	(100 × 100)/200 = 50
Lower middle	(60 × 100)/200 = 30
Lower right	(40 × 100)/200 = 20
	Σ = 200

After calculating expected frequency for each cell, the chi-square may be calculated as shown below:

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
30	25	+5	25	1
15	15	0	0	0
5	10	–5	25	2.5
25	25	0	0	0
10	15	–5	25	1.67
15	10	+5	25	2.5
45	50	–5	25	0.5
35	30	+5	25	0.83
20	20	0	0	0
$\Sigma f_o = 200$	$\Sigma f_e = 200$	$\Sigma f_o - \Sigma f_e = 0$		$\Sigma = 9.00$

$$df = (r - 1)(K - 1) = (3 - 1)(3 - 1) = 2 \times 2 = 4$$

Entering the probability table of chi-square, we find that the value of chi-square for $df = 4$ at the 0.05 level should be 9.488. As the obtained chi-square is below it $(P > 0.05)$, we conclude that the null hypothesis is retained. Hence, the two variables, namely, educational qualification and educational attainment in the present study are found to be independent. For calculating df in a chi-square test the formula, as noted above, is $(r-1)(k-1)$ where r = the number of rows and k is the number of columns.

By applying formula 23.38, the process of computing X^2 becomes much simpler as it avoids many of the arithmetical calculations. Thus we find X^2 in a simpler way the data of 23.23 as under:

$$X^2 = \frac{(30)^2}{25} + \frac{(25)^2}{25} + \frac{(45)^2}{50} + \frac{(15)^2}{15} + \frac{(10)^2}{15} + \frac{(35)^2}{30} + \frac{(5)^2}{10} + \frac{(15)^2}{10} + \frac{(20)^2}{20} - 200$$

$$= (36 + 25 + 40.5 + 15 + 6.67 + 40.83 + 2.5 + 22.5 + 20) - 200$$

$$= 209 - 200 = 9$$

Thus, we get exactly the same value of chi-square as it was obtained with the help of formula 23.37.

Chi-square in 2 × 2 table: When data have been arranged in a 2 × 2 contingency table (where $df = 1$), we need not calculate expected frequency in the manner described above. In such a situation, the chi-square can be directly calculated with the help of the following equation:

$$X^2 = \frac{N[|AD - BC|]^2}{(A+B)(C+D)(A+C)(B+D)} \qquad (23.39)$$

where A, B, C and D = symbols for frequency of four cells in a 2 × 2 table; N = total number of frequencies; bars (||) indicate that in subtracting BC from AD, the sign is ignored.

Suppose the researcher wants to know whether or not the two given items in the test are independent. Both items have been answered in "Yes" or "No" form. The test was administered to a sample of 400 students and the obtained data were as follows:

Table 23.24 Chi-square in a 2 × 2 table

		Item No. 6		
		Yes	No	
Item No. 10	No	180 A	120 B	300
	Yes	90 C	10 D	100
		270	130	400

According to the formula:

$$X^2 = \frac{400[|(180(10) - (120)(90)|]^2}{(300)(100)(270)(130)}$$

$$= \frac{400 \times 81000000}{1053000000} = \frac{32400000000}{1053000000} = 30.769$$

$$df = (r-1)(k-1) = (2-1)(2-1) = 1$$

Entering the probability table of chi-square, we find that for $df = 1$ the value of chi-square at the 0.001 level should be 10.827. As the obtained value of the chi-square is much above it, we conclude that item nos. 6 and 10 are not independent, that is, they are related.

Sometimes it happens that with 1 df, any one of the expected cell frequencies becomes less than 5. In such a situation, a correction called *Yates' correction for continuity* is applied. Some writers have suggested that Yates' correction for continuity should be applied when any of the expected frequencies goes below 10. Where frequencies are large, this correction makes no difference but where frequencies are small, Yates' correction is significant. Yates' correction consists in reducing the absolute value of difference between f_o and f_e by 0.5, that is, each f_o which is larger than f_e is decreased by 0.5 and each f_o which is smaller than f_e is increased by 0.5. The formula for chi-square in such a situation is as given below.

$$X^2 = \frac{N\left[|AD - BC| - \frac{N}{2}\right]^2}{(A+B)(C+D)(A+C)(B+D)} \qquad (23.40)$$

where subscripts are defined as usual.

Suppose, 60 students (50 boys and 10 girls) were administered an attitude scale. The items were to be answered in "Yes" and "No" form. Their frequencies towards item no. 10 are presented in Table 23.25. The question is: Do the opinions of boys and girls differ significantly?

Table 23.25 Chi-square with Yates' correction in a 2 × 2 table

	Yes	No	
Boys	20 A	30 B	50
Girls	3 C	7 D	10
	23	37	60

According to Equation 23.40,

$$X^2 = 60\frac{\left[|(20)(7) - (30)(3)| - \frac{60}{2}\right]^2}{(50)(10)(23)(37)}$$

$$= \frac{60[|140 - 90| - 30]^2}{425500} = \frac{60 \times 400}{425500} = \frac{24000}{425500} = 0.056$$

In the above example, the expected frequency (23× 10/60) is less than 5. Hence, chi-square has been calculated by Equation 23.40. Entering the table for chi-square, we find that for $df = 1$ the value of chi-square at the 0.05 level should be 3.841. Since the obtained value is less than it $(p >> 0.05)$, we conclude that the opinions of boys and girls do not differ significantly.

There are some *assumptions* of chi-square. Some important assumptions and restrictions for using chi-square test are presented below.

(a) *Random sampling:* For the proper use of chi-square test, it is assumed that the sample under study is selected randomly from the population (Gravetter & Wallnau 1985).

(b) *Independence of observations:* By independence of observations, it is meant that each observed frequency is generated by a different subject or a person. This should not be confused with the concept of independence between variables as found in the test of independence. A chi-square test would be misleading and inappropriate if a subject could produce responses that can be classified in more than one category or contribute to more than one frequency count to any single category. In simple words, chi-square test should not be used if the scores are based on the same people being tested more than once (Aron, Aron & Coups 2006).

(c) *Size of expected frequencies:* The size of the expected frequency (f_e) tends to distort the value of chi-square. Let us consider the computation of chi-square for a single cell. Suppose a cell has the $f_e = 1$ and $f_0 = 6$. The contribution of this cell to the total chi-square will be:

$$\text{Cell}(c) = \frac{(f_o - f_e)^2}{f_e} = \frac{(6-1)^2}{1} = \frac{(5)^2}{1} = 25$$

Now consider another instance where $f_o = 15$ and $f_e = 10$. The difference between f_o and f_e is still 5 but the contribution of this cell to the total chi-square value differs from that of the first case.

$$c = \frac{(f_o - f_e)^2}{f_e} = \frac{(15-10)^2}{10} = \frac{(5)^2}{10} = \frac{25}{10} = 2.5$$

It is clear that a small f_e value can have a great impact upon chi-square value. This problem becomes serious when f_e values are less than 5. In fact, the test is too sensitive when f_e values are extremely small. Thus, the chi-square test should not be used when any of the expected cell frequency is less than 5.

(d) *Assumption of continuity:* For the theoretical distributions, chi-square values are assumed to be a continued variable. This assumption is clearly violated when there is only one degree of freedom. When chi-square is calculated for all possible samples of a given size, the resulting distribution of sample chi-square is not continuous. In such a situation, statisticians have suggested the use of *Yates' correction for continuity*, which consists of subtracting 0.5 from the absolute value of the difference between f_o and f_e (Gravetter & Wallnau 1987).

Associated with this assumption of continuity, there is controversy regarding minimum expected frequency. A research article by Lewis and Burke (1949) has demonstrated nine common errors in the use of chi-square. One error is related to the expected frequencies that are too low. Most statisticians are of view that every cell of contingency table should have a reasonable size of expected frequency. Some of them recommended a minimum of 10, with 5 as bottom limit. Fisher (1938) recommended 10 as minimum. Still others recommended that the minimum should be some proportion or percentage of the total or that it depended on whether the expected frequency were equal or not. A rule that has generally been adopted in the 1 degree of freedom situation is that the expected frequency should be equal to or greater than 5. When $df > 1$, the expected frequency should be equal to or greater than 5 in at least 80% of the cells. When these requirements are not met, other statistical tests are available (Siegel & Castellan 1988). However, a major and significant review of the research on the topic was done by Delucchi (1983). He drew two major conclusions as under:

(a) The chi-square test may be properly used in cases where the expected frequencies are much lower than the value considered permissible. The most important principle seems to be that there should be at least five times as many individuals as there are cells. For example, any cell with a very low expected frequency would be acceptable in 2 ×2 contingency table if there are at least 20 subjects in the study overall.

In case the researcher has a table larger than 2 ×2 with a cell having an extremely low expected frequency as well as the number of subjects or participants are also small, one common step would be to combine related categories to increase expected frequency and reduce the total number of cells. But this is a solution of the last resort. The best solution will be to add more participants to the study.

(b) Even though using chi-square with small expected frequencies may be acceptable, it is still not a wiser step. This is because the probability of getting a significant result, even if the research hypothesis is true, may be quite slim. In other words, with small expected frequencies, the power of chi-square test is very low. Thus the risk of Type II error becomes high.

Chi-square test has some *limitations*, too. Important limitations are as under:

(a) As said earlier, chi-square test can not be used when the researcher has counted or included some people more than once. This error produces what is called as an *inflated N* and is extremely serious and may easily lead to the rejection of the null hypothesis when, in fact, it is true (Type I error).

(b) Another limitation of chi-square stems from the fact that the value of chi-square is *proportional* to the sample size. Let us take an example. In a 2 ×2 table the following data were obtained with respect to relationship between sex and the level of anxiety.

		Sex		
		Male	Female	
Level of anxiety	High	32 (27.5)	28 (32.5)	60
	Low	23 (27.5)	37 (32.5)	60
		Σ = 55	65	120

$X^2 = 2.36$ (f_e appears in bracket in each cell)

When the observed frequencies of each cell is doubled, that is, 64, 46, 56 and 74, X^2 would be equal to 5.44 rather than 2.36. Despite the fact that the relationship has *not* changed, the researcher would reject rather than accept null hypothesis. Due to this reason, many researchers prefer to avoid chi-square test when dealing with large samples because the results can be very misleading.

(c) Still another limitation arises from a generally adopted rule related to the situation with small *N*'s or when the expected frequency/proportion/percentage among cell is small. The common rule is that with 1 degree of freedom situation, the expected frequency in all cells should be equal to or greater than 5. When $df > 1$, the expected frequency should be equal or greater than 5 in at least 80% of cells. Somehow when these requirements are not met, other statistical tests are available (Siegel & Castellan 1988)

Mann–Whitney *U* Test

The Mann–Whitney *U* Test is a nonparametric substitute for the parametric *t* test. This test was jointly published by H B Mann and D R Whitney in 1947. Needless to say, the Mann–Whitney *U* test is used when the researcher is interested in testing the significance of difference between two independently drawn samples or groups or two separate and uncorrelated groups. For application of the *U* test it is essential that the data have been obtained on ordinal measurement, that is, they must have been obtained in terms of rank. Where the data have been obtained in terms of scores for application of the Mann–Whitney *U* test, it is essential that those scores be converted into rank without much loss of information. It is not necessary for the application of the Mann–Whitney *U* test that both groups must have unequal size. This test can also be applied to groups having equal size.

The rationale behind Mann–Whitney *U* test is that if two groups are really samples from the sample population, then the ratio of the sum of ranks in two groups would be roughly proportional to the ratio of the number of cases in the two groups. However, if the sums of ranks are not so proportional, the investigator rejects the null hypothesis and finally concludes that the two groups under study differ.

When sample sizes are very small, that is, when both N_1 and N_2 are made up of 20 or less than 20 cases, the students are referred to Siegel's (1956) book where methods and probability tables that apply to such samples are given. In this book we shall deal with methods of calculating the Mann–Whitney *U* test, which are concerned with larger sample size, say, more than 20 cases. The reason for selecting only these methods is that in research, we are more often confronted with sample sizes which are greater than 20. The equations for calculating the Mann–Whitney *U* test are as given below.

$$U = N_1 N_2 + \frac{N_1(N_1+1)}{2} - \sum R_1 \qquad (23.41)$$

$$U = N_1 N_2 + \frac{N_2(N_2+1)}{2} - \sum R_2 \qquad (23.42)$$

Let us illustrate the calculation of the Mann–Whitney *U* test, which can be done by either Equation 23.41 or by Equation 23.42. Table 23.26 presents the scores of two groups on the Lie Scale. Group I has 10 subjects and Group II has 21 subjects. The first step is to rank all the scores in one combined distribution in an increasing order of size. In Table 23.26 the lowest score (taking both sets of scores together) is 7 (second column) and hence, we give it a rank of 1. The next score is 8, which is again in the second column and it has been given a rank of 2. The third score from below is 10 (in the first column), which has been given a rank of 3. In this way, ranking is continued until all scores receive ranks. Subsequently, the two columns of ranks are summed. At this juncture, a check on arithmetical calculation is imposed. The check is that the sums of these two columns must be equal to $N(N + 1)/2$.

$$\text{Check: } \sum R_1 + \sum R_2 = \frac{N(N+1)}{2} = 88.5 + 407.5 = 496; \ \frac{(31)(32)}{2} = 496$$

Hence, we can proceed:

$$\text{(by Equation 23.41) } U = N_1N_2 + \frac{N_1(N_1+1)}{2} - \sum R_1 = (10)(21) + \frac{(10)(10+1)}{2} - 88.5 = 176.5$$

$$\text{(by Equation 23.42) } U = N_1N_2 + \frac{N_2(N_2+1)}{2} - \sum R_2 = (10)(21) + \frac{(21)(21+1)}{2} - 407.5 = 33.5$$

Table 23.26 Calculation of the Mann–Whitney *U* test from larger sample sizes

Gr. I ($N_1 = 10$)	Gr. II ($N_2 = 21$)	R_1	R_2
18	32	7	13
14	40	5	18
30	31	11	12
10	39	3	16.5
39	15	16.5	6
26	8	9	2
27	47	10	19
19	33	8	14
35	52	15	22
11	48	4	20
	7		1
	50		21
	61		27
	58		24
	53		23
	59		25
	60		26
	65		30
	63		28
	67		31
	64		29
		$\Sigma R_1 = 88.5$	$\Sigma R_2 = 407.5$

It is the lower value of U test that we want. For testing the significance of the obtained *U*, its value is converted into *z* score as shown below.

$$z = \frac{U - \frac{N_1N_2}{2}}{\sqrt{\frac{(N_1)(N_2)(N_1+N_2+1)}{12}}} \qquad (23.43)$$

$$= \frac{176.5 - \frac{(10)(21)}{2}}{\sqrt{\frac{(10)(21)(10+21+1)}{12}}} = \frac{71.5}{23.664} = 3.02$$

A *z* score from ±1.96 to ±2.58 is taken to be significant at the 0.05 level of significance and if the *z* score is greater than even ±2.58, we take it to be significant at the 0.01 level. Since the obtained *z* is 3.02, we can take the value of the Mann–Whitney *U* to be a significant one. Rejecting the null hypothesis, it is concluded that the two groups differ significantly on the measures of the Lie Scale.

According to Equation 23.42, the obtained value of the Mann–Whitney *U* differs but the value of the *z* score would remain unaffected. However, the sign of the *z* score would be changed.

$$z = \frac{33.5 - \frac{(10)(21)}{2}}{\sqrt{\frac{(10)(21)(10+21+1)}{12}}} = \frac{-71.5}{23.664} = -3.02$$

Thus we obtain the same *z* score but the sign is negative. The change of sign makes no difference in the interpretation of the Mann–Whitney *U* test.

Rank-Difference Methods

The methods of correlation based upon rank differences are very popular among behavioural scientists. There are two most common methods which are based upon the differences in ranks assigned on the X and Y variables. One is the *Spearman rank-difference method* and the other is the *Kendall rank-difference method*.

The Spearman rank-difference method symbolized by ρ (read as *rho*) is a very popular method of computing the correlation coefficient between two sets of ranks or between two sets of scores converted into ranks. When a researcher is not concerned with whether or not a relation is linear, Spearman correlation is a better choice. The method has been named after Spearman who discovered it. This method is applicable when the number of pairs of scores or ranks is preferably small, that is, 30 or below.

The equation is:

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)} \tag{23.44}$$

where ρ = Spearman's rank-difference correlation coefficient; D = difference between rank_1 and rank_2; and N = number of pairs of ranks or scores.

To illustrate the calculation of ρ let us consider the data given in Table 23.27, which show the scores of 12 students on the intelligence test (X) as well as on the educational test (Y). By computing ρ by Equation 23.44, the first step is to rank both sets of scores separately giving the highest score a rank of 1, the next highest score a rank of 2, and so on. Then, keeping the algebraic signs in view, the difference between two sets of ranks is computed. This is noted under column D. Subsequently, each difference is squared and noted under column D^2. Substituting the values in the equation, we get a ρ of –0.185. Following Siegel & Castellan (1988, Table Q, p 360) we can test the significance of the obtained ρ. Since the obtained value of ρ is less than the value given at the 0.05 level ($\rho > 0.05$) for $N = 12$, we can accept the null hypothesis and can conclude that X and Y are independent and whatever correlation has been found is due to the chance factor.

Table 23.27 Illustration of the Spearman rank-difference correlation

X	Y	Rank_1	Rank_2	D $(R_1 - R_2)$	D^2
47	68	8.5	1	+ 7.5	56.25
50	60	5.5	2.5	+ 3	9.00
70	54	2	7	– 5	25.00
72	53	1	8	– 7	49.00
46	60	10	2.5	– 7.5	56.25
50	55	5.5	6	– 0.5	0.25
42	48	11	9	+ 2	4.00
58	30	3	12	– 9	81.00
55	45	4	10	– 6	36.00
36	43	12	11	+ 1	1.00
49	59	7	4	+ 3	9.00
47	56	8.5	5	+ 3.5	12.25
				$\Sigma D = 0.0$	$\Sigma D^2 = 339.00$

$$\rho = 1 - \frac{6(339)}{12(144-1)} = 1 - \frac{2034}{1716} = 1 - 1.185 = -0.185$$

Relation between Spearman Correlation and Pearson Correlation

If we compute Pearson correlation and Spearman correlation from the same set of data, how will the two correlations seem to be related? Since these two methods of correlations are measuring different things, they will be producing different values. As we know, Pearson correlation measures linear relationship whereas Spearman correlation measures monotonic relationship.

In most cases, the value obtained by Spearman correlation will be larger (that is, closer to +1.00 or –1.00) than the value obtained by Pearson correlation from the same data. This happens because it is easier to be perfect by Spearman's criteria than by Pearson's criteria. However, there is one situation where Pearson correlation will be producing larger value than Spearman correlation. When there is extreme score in the data, this single extreme score can have larger influence on the value of Pearson correlation. The single extreme score tends to exaggerate the magnitude of the correlation bringing it nearer to 1.00 or –1.00 than would be expected from other scores points. The exaggerated influence of extreme score is eliminated with Spearman correlation because the ranking process reduces the distance between adjacent scores to exactly one (that is, first to second, second to third, third to fourth, and so on). This fact is being illustrated below ($N = 05$).

X	Y
0	2
1	0
2	3
3	1
15	15
$\Sigma = 21$	$\Sigma = 21$

Pearson $r = 0.996$

Spearman *rho* = 0.50

It is obvious from the above example that the single score point (15 – 15) has inflated the value of Pearson r (0.96) making it much larger than Spearman *rho* (0.50) and bringing it closer to +1.00.

Another method of computing the rank-difference correlation has been developed by Kendall. The method is known as Kendall's *tau* for which the formula is as follows:

$$T = \frac{S}{(1/2)N(N-1)} \tag{23.45}$$

where T = Kendall's tau, S = actual total; and N = number of objects or scores which have been ranked.

Table 23.28 Scores of 12 students on X and Y test

	A	B	C	D	E	F	G	H	I	J	K	L
X	20	26	17	16	15	23	22	24	19	28	30	10
Y	70	80	40	45	38	49	77	76	72	47	36	35

Suppose 12 students have been administered two tests and their scores are presented in Table 23.28. The first step is to rank both sets of scores giving the highest score a rank of 1, the next higher a rank of 2, and so on. Table 23.29 presents the ranks based upon two sets of the scores given in Table 23.28. Subsequently, the ranks of the *X* test are rearranged in a way that they appear in a natural order like 1, 2, 3... .

Table 23.29 Ranks based upon two sets of scores given in Table 23.28

	A	B	C	D	E	F	G	H	I	J	K	L
X	7	3	9	10	11	5	6	4	8	2	1	12
Y	5	1	9	8	10	6	2	3	4	7	11	12

Accordingly, ranks on the *Y* test are adjusted. Table 23.30 presents the ranks in a rearranged order. Subsequently, the value of *S* is computed. For this, we start with the rank on the *Y* test from the left side. The first rank on the left side is 11. Count the number of ranks which are above 11 and the number of ranks which are below 11 separately.

Table 23.30 Rearranged order of ranks

	A	B	C	D	E	F	G	H	I	J	K	L
X	1	2	3	4	5	6	7	8	9	10	11	12
Y	11	7	1	3	6	2	5	4	9	8	10	12

Only one rank (that is, 12) falling at the right of the first rank on the *Y* test is above 11 and the remaining 10 ranks fall below it. Hence, its contribution to *S* would be equal to 1 – 10. Likewise, the second rank on the *Y* test is 7. The four ranks falling right of 7, are above 7 and 6 ranks are below it. Hence, its contribution to *S* would be 4 – 6. Identical procedures are repeated for other ranks on the *Y* test. Thus:

$$S = (1-10) + (4-6) + (9-0) + (7-1) + (4-3) + (6-0) + (4-1) + (4-0) + (2-1) + (2-0) + (1-0)$$

$$= (-9) + (-2) + (9) + (6) + (1) + (6) + (3) + (4) + (1) + (2) + (1) = 33 - 11 = 22.$$

Substituting in the form of formula:

$$T = \frac{S}{\frac{1}{2}N(N-1)} = \frac{22}{\frac{1}{2}12(12-1)} = \frac{22}{66} = 0.333$$

The significance of *tau* is tested by converting it into a *z* score, the formula for which is as given below.

$$z = \frac{T}{\sqrt{\frac{2(2N+5)}{9N(N-1)}}} \tag{23.46}$$

Hence,

$$z = \frac{0.33}{\sqrt{\frac{2[(2)(12)+5)]}{9(12)(12-1)}}} = \frac{0.33}{\sqrt{0.0488}} = \frac{0.33}{0.2209} = 1.4938$$

Since the obtained *z* score is less than 1.96, we can say that this is not significant even at the 0.05 level. Accepting the null hypothesis, we can say that the given set of scores is not correlated. According to Siegel (1956, 214), *tau* has one advantage over *rho*, and that is that the former can be generalized to partial correlation. If both *tau* and *rho* are computed from the same data, the answer will not be the same and hence, numerically, they are not equal.

Kendall's Partial Rank-order correlation

Whenever correlation exists between two variables, there is much possibility that the correlation is due to the association between each of these two variables and a third variable. For example, among a group of schoolchildren of diverse ages, one might find a high correlation between height and weight. This correlation may not reflect a genuine correlation between height and weight, rather it may result from the fact that both height and weight are associated with a third variable, that is, age. Likewise if one finds a high correlation between size of vocabulary and memorization ability among a group of children of the same age, this correlation may not be true due to the fact that both these variables, that is, size of vocabulary and memorization ability may be associated with a third variable, that is, intelligence.

In designing an experiment, the researcher has the alternative of either introducing experimental control to eliminate or control the impact of third variable by using some statistical methods to eliminate the influence of third variable. Using experimental control, the researcher may choose children of the same intelligence level and then correlate the two variables, that is, size of vocabulary and memorization ability. If somehow the experimental control is not feasible, the statistical methods may be applied for exercising such control. Partial correlation is one such method where the effect of variation in a third variable upon the relation between other two variables, namely, *X* and *Y* are eliminated. Thus by using partial correlation in the above example, the researcher can compute the relation between memorization ability and the size of vocabulary by holding the influence of intelligence constant.

Kendall's partial rank-order is one nonparametric statistical method where the researcher correlates the two variables measured on ordinal scale or converted into ordinal measurement by holding the influence of the third variable constant. Here no assumptions about shape of population distribution of scores need to be made. When *X* and *Y* are to be correlated while controlling the *Z* variable, it is worthwhile that the Kendall's *tau* should be calculated between *X* and *Y*, *X* and *Z* and *Y* and *Z*. The formula for calculating Kendall's partial rank-order correlation coefficient is as under.

$$T_{xy.z} = \frac{T_{xy} - (T_{xz})(T_{yz})}{\sqrt{(1-T_{xz}^2)(1-T_{yz}^2)}} \tag{23.47}$$

where, $T_{xy.z}$ = Kendall's partial rank order correlation
T_{xy} = Kendall's *tau* between *X* and *Y*
T_{yz} = Kendall's *tau* between *Y* and *Z*
T_{xz} = Kendall's *tau* between *X* and *Z*.

Let us take an example. Suppose the size of vocabulary (*X*), memorization ability (*Y*) and intelligence (*Z*) are three variables. The researcher wants to compute $T_{xy.z}$, that is, he wants to have partial rank-order correlation between *X* and *Y* by controlling the impact of *Z*. Suppose further that the values of Kendall's *tau* obtained from a distribution (*N* = 22) are as under:

$T_{xy} = 0.55$
$T_{xz} = 0.46$ (Kendall's rank-order correlation)
$T_{yz} = 0.63$

Now, substituting the values in formula 23.47, we get:

$$T_{xy.z} = \frac{0.55-(0.46)(0.63)}{\sqrt{(1-0.42^2)(1-0.63^2)}}$$

$$= \frac{0.26}{0.70} = 0.37$$

Thus when the impact of intelligence (*Z*) is controlled or partialed out, the correlation between *X* and *Y* is corrected or lowered from 0.55 to 0.37.

With larger *N* (exceeding 20), the significance of $T_{xy.z}$ may be tested by calculating the value of *z* as under:

$$z = \frac{3T_{xy.z}\sqrt{N(N-1)}}{\sqrt{2(2N+5)}} \tag{23.48}$$

Thus,

$$z = \frac{(3)(.37)\sqrt{22(22-1)}}{\sqrt{2(2\times 22+5)}}$$

$$= \frac{23.85}{9.90} = 2.41$$

Since obtained value of *z* exceeds 1.96 but falls short of 2.58, it may be concluded that $T_{xy.z}$ is significant at 0.05 level but not at 0.01 level.

Coefficient of Concordance *W*

The coefficient of concordance symbolized by the letter *W* has been developed by Kendall and is a measure of correlation between more than two sets of ranks. Thus, *W* is a measure of correlation always among more than two sets of rankings of events, objects and individuals. When the investigator is interested in knowing the inter-test reliability, *W* is chosen as the most appropriate statistic. One characteristic of *W* which distinguishes it from other methods of correlation is that it is either zero or positive. It cannot be negative. *W* can be computed with the help of the formula given below.

$$W = \frac{S}{\frac{1}{12}K^2(N^3-N)} \tag{23.49}$$

where *W* = coefficient of concordance; *S* = sum of squares of deviations from the mean of R_j; *K* = number of judges or sets of rankings; and *N* = number of objects or individuals which have been ranked.

Suppose four teachers (*A*, *B*, *C* and *D*) ranked 8 students on the basis of performance shown in the classroom. The ranks given by the four teachers are presented in Table 23.31. The details of the calculations have also been shown.

Table 23.3. Ranks given by four teachers to eight students on the basis of classroom performance

Teachers	Students							
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
A	3	4	7	5	8	6	2	1
B	2	3	6	4	8	7	1	5
C	1	3	5	6	8	7	2	4
D	3	4	2	5	7	6	1	8
R_j	9	14	20	20	31	26	6	18

$$\text{Mean of } R_j = \frac{9+14+20+20+31+26+6+18}{8} = \frac{144}{8} = 18$$

$$S = (9-18)^2 + (14-18)^2 + (20-18)^2 + (20-18)^2 + (31-18)^2$$

$$+ (26-18)^2 + (6-18)^2 + (18-18)^2$$

$$= (-9)^2 + (-4)^2 + (2)^2 + (2)^2 + (13)^2 + (8)^2 + (-12)^2 + (0)^2 = 482$$

Now substituting in the form of Equation 22.17,

$$W = \frac{482}{\frac{1}{12}(4)^2(8^3-8)} = \frac{482}{672} = 0.717 = 0.72$$

When N > 7, the significance of *W* is tested by converting its value into X^2 with the help of the following equation:

$$X^2 = K(N-1)W \tag{23.50}$$

Thus $X^2 = 4(8-1)(0.72) = 20.16$ and *df* in this situation is always equal to $N-1$. Hence $df = 8-1 = 7$. Entering in the probability table for chi-square, we find that the value of chi-square for $df = 7$ at 0.05 level of significance should be 18.475. Since the obtained value of the chi-square exceeds this required value, we can take this value of *W* as a significant one. Thus, rejecting the null hypothesis we can say that there is an overall significant relationship in ranking done by the four teachers.

Median Test

The median test is used to see if two groups (not necessarily of the same size) come from the same population or from populations having the same median. A median test may be readily extended to more than two samples, that is for several independent groups (Kurtz & Mayo 1980). In the median test, the null hypothesis is that there is no difference between the two sets of scores because they have been taken from the same population. If the null hypothesis is true, half of the scores in both groups should lie above the median and the remaining half of the scores should lie below the median. Table 23.32 presents the scores of two groups of students in an arithmetic test. The first step in computation of a median test is to compute a common median for both distributions taken together.

Table 23.32 Scores of 30 students in an arithmetic test

Gr. A (N = 16)	16, 17, 8, 12, 14, 9, 7, 5, 20, 22, 4, 26, 27, 5, 10, 19
Gr. B (N = 14)	28, 30, 33, 40, 45, 47, 40, 38, 42, 50, 20, 18, 18, 19

For computing the common median, both the distributions are pulled together as shown in the next page.

Scores	f
49–53	1
44–48	2
39–43	3
34–38	1
29–33	2
24–28	3
19–23	5
14–18	5
9–13	3
4–8	5
	$N = 30$

$$\text{Median} = l + \frac{(N/2 - F)i}{f_m}$$

$$= 18.5 + \frac{(30/2 - 13)5}{5}$$
$$= 20.5$$

Subsequently, a 2 × 2 contingency table is set as follows.

Now, the chi-square test can be applied. For computing chi-square from a 2 × 2 table, we may follow Equation 23.39. Yates' correction is not needed here because none of the cells contains an expected frequency less than 5.

	Above Mdn	Not above Mdn	
Gr. A	A 3	B 13	16
Gr. B	C 10	D 4	14
	13	17	30

Now substituting the values in Equation 23.39, we get:

$$X^2 = \frac{30[\,|(3)(4) - (13)(10)|\,]^2}{(16)(14)(13)(17)}$$

$$= \frac{30[\,|12 - 130|\,]^2}{49504} = \frac{417720}{49504} = 8.438 = 8.44$$

$$df = (r-1)(c-1) = (2-1)(2-1) = 1$$

Entering the probability table for chi-square we find that for $df = 1$, the chi-square value at the 0.01 level should be 6.635. Since the obtained value of the chi-square exceeds this value ($p < 0.01$), we can reject the null hypothesis and conclude that the two samples have not been drawn from the same population or from populations having same median.

Kruskal–Wallis *H* Test

The primary difference between the *F* test and the Kruskal–Wallis *H* test on the one hand and the Friedman test on the other (to be considered in the next section) is that the former is a parametric analysis of variance, whereas the *H* test and the Friedman test are the nonparametric analysis of variance. The *H* test is a one-way nonparametric analysis of variance and the Friedman test is a two-way nonparametric analysis of variance. The *H* test is used when the investigator is interested in knowing whether or not groups of independent samples have been drawn from the same population. If the obtained data do not fulfil the two basic parametric assumptions, namely, assumption of normality and assumption of homogeneity of variances, the *H* test is the most appropriate statistic. The equation for the *H* test is as given below.

$$H = \frac{12}{N(N+1)}\left[\sum \frac{R_j^2}{N_j}\right] - 3(N+1) \qquad (23.51)$$

where N = number in all samples combined; R_j = sum of ranks in j sample; and N_j = number in j sample.

Table 23.33 *H* test from scores obtained by three groups on Lie Scale

Gr. A	$(N = 6)$	Gr. B	$(N = 8)$	Gr. C	$(N = 10)$
15	(14)	17	(15)	20	(17)
10	(9)	9	(8)	25	(19)
8	(6)	8	(6)	13	(12)
5	(3)	14	(13)	11	(10)
6	(4)	2	(1)	26	(20)
4	(2)	8	(6)	24	(18)
		12	(11	36	(24)
		18	(16)	30	(23)
				29	(22)
				27	(21)
$R_j = 38$		$R_j = 76$			$R_j = 186$

Data to illustrate the calculation of the *H* test have been given in Table 23.33. Three groups of students were administered a Lie Scale and their scores are presented in Table 23.33. The first step is to combine all the scores from all of the groups and rank them with the lowest score receiving a rank of 1 and the largest score by rank *N*. Ties are treated in the usual fashion in ranking. Subsequently, the sum of ranks in each group or column is found and the *H* test determines where these sums of ranks are so disparate that the three groups cannot be regarded as being drawn from the same population. The ranks assigned to each score earned by the member of the group are given In brackets. Now, substituting the value in Equation 23.51, we have

$$H = \frac{12}{(24)(25)}\left[\sum \frac{(38)^2}{6} + \frac{(76)^2}{8} + \frac{(186)^2}{10}\right] - 3(24+1)$$

$$= \frac{53067.192}{600} - 75 = 88.445 - 75 = 13.445 = 13.44$$

When each sample has six or more than six cases, the H test is interpreted as chi-square. In such a situation, df = number of group or samples minus one. So, here $df = 2$. Entering the probability table for chi-square we find that for $df = 2$, the value of chi-square at the 0.01 level of significance should be 9.210. Since the obtained value of H test exceeds this required value, it can be said that the H value is significant. Rejecting the null hypothesis, we conclude that the samples are independent and they have not been drawn from the same population.

When the number of cases in samples is from one to five, the H test can be interpreted through special tables, which can be found in Siegel (1956) and Downie & Heath (1970).

Friedman Test

As mentioned above, the Friedman test is a two-way nonparametric analysis of variance. When the groups are matched and doubts exist about the two basic parametric assumptions, namely, the assumption of normality and the assumption of homogeneity of variances, the investigator resorts to the Friedman test for testing whether or not the samples have been drawn from the same population. Data to illustrate the calculation of the Friedman test have been presented in Table 23.34. The equation for calculating Friedman test is as follows:

$$X_r^2 = \frac{12}{NK(K+1)}(R_j)^2 - 3N(K+1) \qquad (23.52)$$

where X_r^2 = Friedman test; N = number of rows; K = number of columns; and R_j = separate sums of ranks of each column.

Table 23.34 Friedman test of two-way ANOVA from scores of three groups on a recall test

	I	II	III	IV	V
Gr. A	12 (4)	8 (2)	4 (1)	10 (3)	16 (5)
Gr. B	10 (3)	7 (2)	6 (1)	11 (4)	17 (5)
Gr. C	7 (2)	8 (3)	4 (1)	12 (5)	10 (4)
R_j	9	7	3	12	14

The first step in calculation of the Friedman test is to rank each score in each row separately, giving the lowest score in each row a rank of 1 and the next lowest score in each row a rank of 2, and so on. The ranking can also be done in reversed order, that is, giving the highest score in each row a rank of 1, the next highest score in each row a rank of 2, and so on. The rank assigned to each score in each row is given in parentheses. The Friedman test is applied to determine whether or not the rank totals symbolized by R_j differs significantly. Now, substituting the values in Equation 23.52, we have

$$X_r^2 = \frac{12}{(3)(5)(5+1)}\left[(9)^2 + (7)^2 + (3)^2 + (12)^2 + (14)^2\right] - 3(3)(5+1)$$

$$= 63.866 - 54 = 9.87$$

When the number of rows (N) and the number of columns (K) are too small, the significance of the Friedman test can be ascertained with the help of special tables (Siegel 1956). For example, when $K = 4$, $N = 2$ to 4 or when $K = 3$, $N = 2$ to 9, the significance of the Friedman test can be done through these special tables. But when the number of rows and columns are greater than those said above, the Friedman test is interpreted as the chi-square test. In the present example, the significance of the Friedman test would be interpreted in terms of chi-square. The df is always equal to $K - 1$ for chi-square applied as a test of significance of the Friedman test. Hence df in the present example would be $K - 1 = 5 - 1 = 4$. Entering the table for df 4, we find that the chi-square should be 9.488 at the 0.05 level of significance. Since the obtained value of the Friedman test exceeds this value ($p < 0.05$), we reject the null hypothesis and conclude that the three matched groups differ significantly.

CORRELATION AND REGRESSION

Correlational methods are the most commonly used statistical techniques in the testing field. We have already discussed some important methods of correlation, namely, Pearson r, Spearman rank difference method and Kendall's *tau*.

As we know, a correlation coefficient is a mathematical index that describes the direction and magnitude of a relationship. There is also a related technique called *regression*, a term first used by Sir Francis Galton in 1885, and today it is used to make a prediction about scores on one variable on the basis of known score on another variable or possibly several other variables (Ferguson & Takane 1989). In fact, these predictions are done from the *regression line*, which is defined as the best-fitting straight line through a set of points in a scatter diagram. It is estimated by using the *principle of least square* which, in fact, minimizes the squared deviations around the regression line. Let us explain through an example.

As discussed above, the mean is the point of least squares for any single variable. In other words, the sum of squared deviations around the mean will be less than it is around any value other than mean. For example, the mean of the five scores, namely, 2, 3, 4, 5 and 6 is $\Sigma X/N = 20/5 = 4$. The squared deviation of each score around the mean can now be easily determined. For score 6, the squared deviation is $(6-4)^2 = 4$; for score 5, the squared deviation is $(5-4)^2 = 1$. The score 4 is equal to the mean and therefore, the squared deviation around mean will be $(4-4)^2 = 0$. Thus, by definition the mean will always be the point of least squares. The regression line is the line of least squares or running mean in two dimensions or in the space created by two variables.

As narrated above, a regression line is the best-fitting straight line through a set of several points in a scatter diagram, which is the picture of the relationship between two variables. The regression line is described as a mathematical index called *regression equation*. The regression equation is an equation for predicting the most probable value of one variable from known value of another. The general linear regression equation for the straight line is:

$$Y = bx + a \qquad (23.53$$

where, a is the *intercept*, that is, value of Y when X is zero. In other words, the point at which the regression line crosses the Y-axis (a) is found by using the following formula:

$$a = Y - bx$$

where, b is the slope of regression line or it is called the regression coefficient. It is expressed as the *ratio* of sum of squares for the covariance to the sum of squares for X. In other words, $b = \frac{\text{vertical change}}{\text{horizontal change}}$. In Figure 23.8, two regression lines have been drawn based on equation $Y = a + bx$.

The regression equation gives a predicted value for Y, on the basis of X. This predicted value is called Y'. The variable *being predicted* is called the *criterion variable* and the variable from which it is

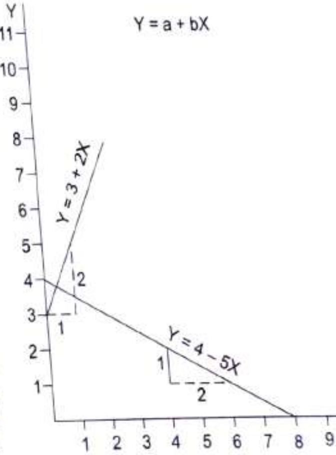

Fig. 23.8

being predicted is called as *predictor variable*. In equations, the predictor variable is usually labeled as X and the criterion variable as Y. Thus X predicts Y. In the above formula, Y is the person's predicted score on the criterion variable and X is the person's score on the predictor variable. In fact, the actual score and predicted score on Y is *rarely* exactly the same. Some difference between the observed and predicted score, that is, $Y - Y'$ occurs and this is technically called the *residual*. Symbolically, the residual thus, is defined as $Y - Y'$. The best-fitting line actually keeps the residual to minimun. In other words, it minimizes the deviation between observed and predicted Y score. Since residuals can be either positive or negative and will cancel to zero if averaged, the best-fitting line is rightly estimated by squaring each residual. In this way, it can be said that the best-fitting line is best obtained by keeping these squared residuals as small as possible

What is the difference between correlation and regression? In fact, there is a difference between these two and the difference is that the correlation coefficient has a reciprocal nature whereas regression lacks this property. The correlation between X and Y will always be the same as the correlation between Y and X. For example, if the correlation between test anxiety and test performance is .64, the correlation between test performance and test anxiety is also .64. We cannot claim a similar thing about regression. In fact, regression is used to transform one variable into estimated scores on the other variable. In other words, regression is often used to predict raw scores on Y on the basis of raw scores on X or vice versa. For example, through regression equation the investigator might be interested in predicting score on the intelligence test on the basis of score on, say, the anxiety test. The coefficient denoting regression of X on Y is usually not the same as the coefficient denoting the regression of Y on X. Since regression uses the raw scores of the variables, the trait of reciprocity does not hold true.

Making predictions on the basis of correlational research is referred to as regression analysis. Regression analysis tends to show how changes in one variable are related to changes in another variable. In psychological testing, regression is often used to determine whether changes in test scores are related to changes in performance. For example, do persons who score higher on mechanical reasoning test perform better as a mechanical engineer? In fact, regression analysis and related correlational methods show the extent to which these variables are linearly related. Besides, they also often create an equation that estimates scores on the criterion (such as performance as mechanical engineer) on the basis of scores on a predictor (such as scores on mechanical reasoning test).

Why is regression also called *prediction?* As we know, regression literally means going back or returning. The term regression is used here because the predicted scores on the criterion variable is closer to the mean of the criterion variable as compared to the distance from the value of the predictor variable to the mean of the predictor variable. However, this rule does not apply when there is perfect correlation between the criterion variable (Y) and the predictor variable (X). Thus we can think in terms of predicted value of the criterion variable regressing or going back toward the mean of the criterion variable.

MAJOR TERMS AND ISSUES IN CORRELATION AND REGRESSION

When we go through the report of correlational analysis, we often need to know about the meaning of some important terms and related issues. Such important terms are: *residual, coefficient of determination, coefficient of alienation, standard error of estimate, shrinkage, cross validation, correlation-causation problem* and *third variable explanation*. A brief discussion of these terms follows.

1. Residual

A short form of residual variance is used in statistics, especially in regression analysis and analysis of variance (ANOVA) and it refers to that part of the variance in the dependent variable that is not accounted for by the independent variables and thus, attributed to chance variation. In regression equation, the difference between a score and its expected value according to a theoretical model is called as residual. As we know, a regression equation gives a predicted value of Y' for each

value of X. In addition to those predicted values, there are observed values of Y. The difference between the predicted value and the observed values is known as residual and symbolically, residual is defined as $Y - Y'$. Let's take an example. Suppose a person has an observed score of 10 and predicted score of 8.46. The residual will be:

$$10 - 8.46 = 1.54$$

In regression analysis, residual has some obvious properties.

(a) The sum of residuals is always equal to zero (0). That is $\Sigma(Y - Y') = 0$

(b) According to the principle of least squares, the sum of the squared residuals is the smallest value. In other words, $\Sigma(Y - Y') =$ smallest value.

2. Coefficient of determination (r^2)

The value of r^2 is called as *coefficient of determination* because it measures the proportion of variability in one variable that is *determined* by or associated with the relationship of other variable. For example, a correlation of $r = 0.90\ (-0.90)$ means that $r^2 = (0.90)^2 = 0.81$ (or 81%) of variability of Y is determined by X or 81% of variability of X is determined by Y. To illustrate it further, suppose the investigator finds a correlation of $r = 0.75$ between IQ (X) and salary (Y) of the employees. It means the coefficient of determination will be $(0.75)^2 = 0.56$, indicating the fact that 56% of the enhancement in salary is due to the variation in IQ of the employees.

Let us take another concrete example to illustrate the concept of coefficient of determination. Suppose the investigator has been given the task of predicting the annual income of Mr Rajesh. To make the task of prediction, he is provided with three bits of information.

(a) Mr Rajesh wears a of size-9 shoe.

(b) Mr Rajesh has been working for 3 months.

(c) Mr Rajesh has a monthly income of Rs 6,000.

It should be clear that the first bit of information (shoe size) is of no help because there is no correlation between shoe size and annual income. However, the second bit of information (length of employment) does provide some useful information. Since Mr Rajesh has been in employment for the last 3 months only, he is probably in an entry-level job and is earning a relatively low salary. Although this information may improve the accuracy of the prediction, the amount of accuracy would depend on the strength of correlation between salary and seniority. Suppose the investigator gets a correlation of $r = 0.38$ between salary and seniority. A correlation of 0.38 would mean $r^2 = (0.38)^2 = 0.14$ or 14% variability in salary is determined by seniority. Finally, the third bit of information, that is, monthly income is highly useful because there is a perfect correlation ($r = 1.00$) between the monthly income and annual income. Since there is a perfect correlation between monthly income and annual income, the investigator is now better able to predict with perfect accuracy. Here the coefficient of determination is $r^2 = (1.00)^2 = 1.00$ which means that annual income is completely *determined* by monthly income.

3. Coefficient of alienation (K)

The coefficient of alienation is the measure of *nonassociation* between two variables. Thus we can think of K as a measure of *absence* of relationship between X and Y in the same sense that r measures the presence of relationship. Therefore, when $K = 1.00$, $r = 0.00$ and when $K = 0.00$, $r = 1.00$. With the larger coefficient of alienation there is smaller extent of relationship and therefore, less precise is the forecast from X to Y and vice versa.

K is calculated as $\sqrt{1-r^2}$, where r^2 is the coefficient of determination. In the example of seniority and annual income, the coefficient of alienation is $\sqrt{1-0.38^2} = \sqrt{1-0.14} = \sqrt{0.86} = 0.927$. This means that there is high degree of nonassociation between seniority and annual income.

4. Standard error of estimate

The standard error of estimate is a measure of accuracy of prediction. The task of prediction is most accurate when standard error of estimate is relatively small. As it becomes larger, the prediction becomes less and less accurate.

As we know, a regression equation, by itself, allows us to make predictions but it does not provide any information about the accuracy of predictions. Therefore, for measuring the precision of the regression, it is customary to compute a standard error of estimate, which is defined as the standard deviation of the differences between actual values of the predicted variable and those estimated from the regression equation. More briefly, it is simply the standard deviation of the residuals.

When *Y* is predicted from *X*, the formula for standard error of estimate becomes:

$$S_{y.x} = S_y\sqrt{1-r_{yx}^2} \tag{23.54}$$

When *X* is predicted from *Y*, the formula for the standard error of estimate becomes:

$$S_{x.y} = S_x\sqrt{1-r_{xy}^2} \tag{23.55}$$

5. Shrinkage

Sometimes a regression equation is developed on one group of subjects and then used to predict the performance of another group. Shrinkage refers to the amount of *decrease* observed when a regression equation created for one population is applied to another one. Formulas are available to estimate the amount of shrinkage (Jaccard & Wan 1995; McNemar 1969). Let us take an example. Suppose a regression equation is developed to predict the performance of one group of MBA students in the first two semesters on the basis of MAT scores. Although the proportion of variance in performance in the first two semesters might be fairly high, we can expect to account for a smaller proportion of variance, when this equation is used to predict the performance in the these two semesters of another group of MBA students selected on the basis of CAT scores. This decrease in proportion of variance accounted for is known as shrinkage.

6. Cross validation

In psychological and educational researches, the term cross validation is frequently used. As explained in chapter 6 also, in psychometrics, it means evaluating the validity of a test by administering the test to an independent sample drawn from the same population on which it was originally validated. In regression analysis, it means measurement of predictor variables and comparison of predicted and observed values of the criterion variable on a fresh sample drawn from the same population as the sample from which regression equation was originally derived for determining the stability of predictions. For doing so, a standard error of estimate can be found for the relationship between the values predicted by the equation and the values actually observed.

7. Correlation-causation problem

Just because two variables are correlated, does it imply that one has caused the other? We are constantly bombarded with reports of relationships: Cigarette smoking is related to lungs cancer, alcohol consumption is related to birth defects. Do these relationships mean that cigarette *causes* lungs cancer or alcohol consumption *causes* birth defects? The answer is *no*. Although there may be a causal relationship, the simple existence of correlation does not prove it. Thus mere existence of correlation does not prove causality. However, it might lead to other research designed to establish causal relationship between variables. Therefore, the existence of correlation by itself does not allow the investigator to decide about the direction of casuality: X causing $Y(X \to Y)$ or Y causing $X(Y \to X)$. However, there is a way to deal with directionality problem to some extent.

As we know, research psychologists are generally satisfied with attributing causality between X and Y when they occur together, when X precedes Y in time, when X causing Y makes sense in relation to some theory and when other explanations for their co-occurrence can be ruled out. In such situations, X may cause Y but Y can't cause X. Using a procedure called a *cross-lagged panel correlation*, it is possible to enhance one's confidence about directionality. This technique investigates correlations between variables at several points in time. Therefore, it is a longitudinal design. It is called cross *because* they are across two variables and lagged because there is a lag between two times. Eron et al. (1972) conducted one important study in which cross-lagged panel correlation technique was used. They measured preference for watching violent television programme (A) and peer ratings of aggressiveness (B). 875 school students from the third grade from the rural area of New York state participated in a study in 1960. A modest correlation of 0.21 was found between preference for violent TV programme and aggressiveness. After 10 years, that is, in 1970, 427 students from the same group were reassessed on the these two variables, labeled as *13th graders*. By measuring two variables (A and B) at two points in time (Time 1 and Time 2) yields four different sets of measure: A_1, A_2, B_1 and B_2 and six correlation were calculated. A_1 means score on variable A of Time 1 and B_1 means score on variable B at time 1 and so on.

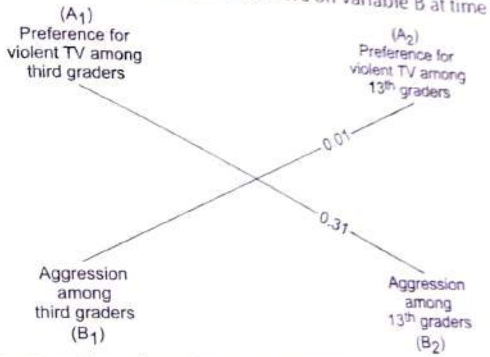

Fig. 23.9.: Cross-lagged panel study showing impact of preference for violent TV programme on later aggression

Of special interest here are the diagonal or cross-lagged correlations because they measure the relationship between variables separated in time (see Figure 23.9). If third-graders aggressiveness really caused a later preference for watching violent TV programme that is, B causing $A(B_1$ vs. $A_2)$, then we should expect a fair-sized correlation between aggressiveness at Time $1(B_1)$ and preference at time $2(A_2)$, but the correlation is virtually zero (+ 0.01). On the other hand, if an early preference for viewing violent TV programmes led to a later aggressiveness, that is A causing $B(A_1$ vs. $B_2)$ then the correlation between preference at Time $1(A_1)$ and aggressiveness at Time $2(B_2)$ should be substantial. This correlation is + 0.31, definitely not large but significant. Based on this finding, it was concluded by Eron et al. (1972) that an early preference for watching violent TV programmes are, at least partially, the cause of later aggressiveness. The other four correlations in the study conducted by Eron et al. (1972) were $0.05(A_1$ vs. $A_2)$, $0.21(A_1$ vs. $B_1)$, $0.38(B_1$ vs. $B_2)$ and $-0.05(A_2$ vs. $B_2)$. These correlations were not of much importance. Thus on the basis of size of the correlation, it is possible to tell which variable causes the other. If the correlation between A_1 and B_2 is greater than that of between B_1 and A_2, one can easily conclude that changes in A causes changes in B rather than vice-versa. Kohle and Berman (1979) also used this cross-lagged panel correlation to answer the question whether attitude causes behaviour o vice versa and reported that attitude was more likely to cause behaviour.

8. Third variable explanation

In correlational research often a high positive correlation between variable A, such as televisio viewing and variable B, such as aggressive behaviour does occur. Does this mean that A causes or B causes A? In other words does excessive TV viewing causes aggressive behaviour or does this mean that aggressive children might prefer to watch a lot of television. There is onother possible explanation for this observed relationship. Thus when a correlation is found between two variables, there is always the possibility that the correlation is due to association between each of the two variables and a third variable. This is called as *third variable explanation*. For example, poor social adjustment can cause both excessive TV viewing as well as aggressiveness. Likewise, among a group of elementary-school children of different ages, the investigator may find a high correlation between size of vocabulary and height. This correlation may not be a genuine relation between these two variables, rather it may result from the fact that both vocabulary size and height are associated with a third variable, that is, age. Thus the apparent relationship between two variables might be the result of some variable not included in the analysis. This external influence is known as third variable explanation.

CHOOSING APPROPRIATE STATISTICAL TESTS

It is customary to choose the appropriate statistical test on the basis of the nature of the obtained data. If the data fulfil the requirement of parametric assumptions, any of the parametric tests which suit the purpose can be selected. On the other hand, if the data do not fulfil the parametric requirements, any of the nonparametric statistical tests which suit the purpose may be selected. Other things to be kept in mind in selection of appropriate statistical tests are the number of independent and dependent variables and the nature of the variables, that is, whether they are nominal, interval or ordinal. When both the independent variable and dependent variable are interval measures and are more than one, multiple correlation is the most appropriate statistic. On the other hand, when they are interval measures and their number is only one, Pearson *r* may be used. As has been noted earlier with ordinal and nominal measures, the nonparametric statistics are the common choice.

Sometimes, researchers transform the measures so that the appropriate statistical test may be applied without loss of much information. For example, if the scores of two groups on interval measures are available but the data do not fulfil the requirement of the *t* test, the researcher can transform the interval measures into ordinal measures and subsequently, apply the Mann–Whitney test.

In this chapter, we have described several statistical tests. These, and few other not described in this text, are summarized in Table 23.35 that shows the purpose of each statistical test and types of data for which it is considered more appropriate.

Table 23.35 Summary of some common statistics with their purposes and appropriateness of data for which they are suitable

Name of statistical test	Parametric (P) / Nonparametric (NP)	Purpose	IV (Independent variable)	DV (Dependent variable)
Pearson *r*	Parametric	Measures linear relationship between two variables	One interval or ratio	One interval or ratio
Multiple correlation (*R*)	Parametric	Measures linear relationship between more than two variables	Two or more intervals or ratios	One interval or ratio
t test	Parametric	Tests the significance of difference between two independent or dependent groups	One nominal	One interval or ratio
ANOVA (*F* test)	Parametric	Tests the significance of difference between more than two groups	One or more nominal variable	One interval or ratio
ANCOVA (Analysis of covariance (*F*)	Parametric	Tests the significance of difference between means of more than two groups while controlling the effect of one or more other variables	One or more nominal variable	One interval or ratio
Multivariate ANOVA* (MANOVA) *F*	Parametric	Tests the significance of mean difference between more than two groups on more than one variable	One or more nominal variable	Two or more intervals or ratio
Chi-square (X^2) test	Nonparametric	Tests whether the proportions of two or more categories differ from the expected proportions	One nominal	One nominal
Median Test	Nonparametric	Tests the difference between median of two independent samples	One nominal	One ordinal
Mann–Whitney *U*-test	Nonparametric	Tests the difference between ranks of two independent samples	One nominal	One ordinal
Phi correlation	Nonparametric	Measures the relationship between two dichotomous variables	One nominal	One nominal
Spearman's rank order *rho* (ρ)	Nonparametric	Measures the relationship between two measures	One ordinal	One ordinal
Kruskal–Wallis *H*-test	Nonparametric	Tests the difference between ranks of more than two independent samples	One nominal	One ordinal
Friedman rank test	Nonparametric	Tests the difference between ranks of more than two observations done on the same sample	One nominal	One ordinal
Coefficient of contingency* (C)	Nonparametric	Tests relation between two variables classified into equal row and column such as 3×3; 4×4, etc.	One nominal	One nominal

---

* Not discussed in the present textbook

Factor Analysis

The technique of factor analysis originated with C Spearman. In fact, factor analysis may be defined as a multivariate statistical method, which is used in the analysis of tables or matrices of correlation coefficient. Spearman analyzed tables of intercorrelations between psychological tests and was able to demonstrate that the intercorrelations could be accounted for in terms of one general factor called as g factor common to all tests, and factors which were specific or unique to each test called specific factor or *S* factor. This was known as the *theory of two factors*. Following Spearman, much work on factor analysis was done by G H Thompson, L L Thurstone, H Hotelling, J P Guilford and R B Cattell.

Factor analysis is basically used to study the interrelationships among a set of variables without reference to a criterion. Thus factor analysis may be understood as a *data-reduction technique*. Broadly, there are two forms of factor analysis: *confirmatory* and *exploratory*. In confirmatory factor analys, is the purpose is to confirm that test scores and variables fit a certain pattern or framework predicted by a theory. For example, if a theory underlying a certain intelligence test predicted that the various subtests belong to three factors such as verbal comprehension, attention, and memory factors, then a confirmatory factor analysis may be undertaken to evaluate the accuracy of this prediction. In fact, confirmatory factor analysis is essential to the validation of many ability tests.

Exploratory factor analysis is comparatively more popular where the central purpose is to summarize the interrelationships among a large number of variables in a concise and accurate way for providing an aid in conceptualization. For instance, an explanatory factor analysis may help a researcher discover that a battery of 25 tests represent only five underlying variables usually called as *factors*.

Factor analysis is usually applied to data where a distinction between dependent variable and independent variable is not meaningful. Major concern is with the description and interpretation of interdependencies within a single set of variables. Factor analysis achieves these purposes in two ways—first, it reduces the original set of variables to a smaller number of variables or factors; and second, factors tend to acquire meaning because of structural properties that may exist within the set of relationships. Thus the process of reducing the number of variables and the concept of structure are basic to the understanding of factor analysis.

All techniques of factor analysis begin with preparation of a complete table of intercorrelations among a set of tests. Such a table is known as *correlation matrix*, which shows the correlation between every variable and every other variables. Subsequently, we find the linear combinations or *principal components* of the variables that describe as many of the interrelationships among the variables as possible. In fact, each principal component, also called as factor, is extracted according to *mathematical rules* (which is complex) that make it independent or uncorrelated with other principal components. The first component is usually most successful in describing the variation among the variables, with each succeeding component somewhat less successful. A factor must be viewed as a variable but with a *simple difference*. Most variables involve direct measurement whereas factors are hypothetical variables derived by a process of analysis from a set of variables obtained by direct measurement. Despite this, a person may be said to have a score on a factor in the same sense that the individual has a score on a test. When factors are uncorrelated, they are said to be *orthogonal* and the method involved herein is called an *orthogonal solution*. When factors are correlated, they are said to be *oblique* and the method involved herein is known as *Oblique solution*. As we know, each factor is a practical combination of the tests used by 'adding in' carefully determined portions of some tests and subtracting out fractions of other tests. What makes factors special is the analytical method used to derive them. Several different methods exist and some of the common ones are: principal component factors, principal axis factors, method of unweighted least squares, image-factoring, alpha factoring, maximum-likelihood methods. The discussion of these methods are beyond the scope of this book.

After correlation matrix, factor matrix is prepared. Factor matrix consists of a table of correlations called factor loading showing the weight or loading of each of the factors of each test. Table 23.36 presents a hypothetical factor matrix involving only two factors. The factors are listed across the top and their weights in each of the 5 tests are given in the appropriate rows.

Table 23.36: A hypothetical factor matrix

Tests	Factor I	Factor II
1. Sentence completion	0.75	–0.24
2. Vocabulary	0.85	0.77
3. Reading comprehension	0.65	–0.60
4. Analogies	0.68	0.67
6. Arithmetical problems	0.43	–0.33

Thus factor loadings range from –1, a perfect negative correlation with a factor, through 0, (no relation to the factor) to +1, a perfect positive correlation. Variables have loadings in each factor but will have high loadings on only one. Normally a variable is considered to contribute meaningfully to a factor only if it has a loadings at least above 0.30 (or below –0.30). Note that the first variable (or test), sentence completion has a strong positive loading of 0.75 on factor I, indicating that this test is reasonably a good index of factor I. Also note that this variable has a modest loadings of –0.24 on factor II, indicating that, to a slight extent, it measures the opposite of this factor that is, high scores on sentence completion tend to signify low scores on factor II and vice versa. In a sense, a factor is produced by 'adding in' carefully determined portions of some tests and perhaps 'subtracting out' fractions of other tests. What makes factors special is the elegant analytical methods undertaken to derive them.

Several different methods for analyzing a set of variables into common factors are available. As early as 1901, Pearson pointed the way for this type of analysis. Later on, Spearman (1927) developed a precursor of modern factor analysis. In America, Kelley (1935) and Thurstone (1947) and in England, Burt (1941) did much to advance the method. Although these methods differ in their initial postulates, most of these methods yield similar results. For a detailed study of the specific procedures of factor analysis, the reader is referred to texts by Comrey and Lee (1992) and Loehlin (1992).

There are some basic *issues* in factor analysis that should not be ignored. A review of literature shows that factor analysis is frequently misused and often misunderstood. Some investigators appear to use factor analysis as a kind of diving rod hoping to find gold hidden underneath tons of dirt. But the reality is that there is nothing magical about factor analysis. In fact, neither factor analysis nor any other statistical analysis can rescue data based on trivial, irrelevant or haphazard measures. However, factor analysis can yield meaningful results only when the research was meaningful to start with. The following three basic issues must be kept in mind:

(i) A particular kind of factor can emerge from factor analysis only when the tests and measures contain that factor in the first place. For example, a memory factor can't emerge from a battery of ability tests if none of the tests requires memory function. Thus, in factor analysis quality of the output depends upon the quality of input.

(ii) For a stable factor analysis, the size of the sample is also considered important. According to Tabachnick and Fidell (2006) it is considered comforting, in general, to have *at least* five subjects for each test or variable. Comrey's guidelines suggest that a sample size of 50 is very poor, 100 poor, 200 fair, 300 good, 500 very good and 1000 excellent for a dependable factor analysis (Gregory 2006)

(iii) A third issue is that we can't overemphasize the extent to which the technique of factor analysis is guided by the subjective choices and theoretical prejudices. A crucial point in this regard is the choice between *orthogonal axes* and *oblique axes*. With orthogonal axes the factors lie at right angles to one another, which means that they are uncorrelated. However, in many cases of clusters of factor loadings, oblique axes provide a better look. With oblique axes, the factors are correlated among themselves. In general, oblique axes are considered better than orthogonal axes because with oblique rotations, it is possible to factor analyze the factors themselves. Such procedure tends to yield one or more second-order factors. Second-order factors tend to provide support for the hierarchical organization of traits/abilities, etc.

Factor analysis has some advantages and disadvantages. Important *advantages* are as under:

(i) Factor analysis permits reduction of a number of variables by combining two or more variables into a single factor. For example, the researcher may combine, through factor analysis, batting, jumping, ball throwing and weightlifting into a single factor called *general athletic ability*.

(ii) Factor analysis helps the researcher in identifying the groups of interrelated variables to examine how they are related with each other. For example, Carroll used factor analysis in formulating his three-stratum theory of intelligence. He reported that a factor called *broad auditory perception* was related to auditory task ability and another factor called *broad visual perception* was related to visual task capability. He also reported that a *global factor* called *g*-factor (or general intelligence) was related to both these factors, that is, broad auditory perception and broad visual perception. This, then, automatically meant that someone high on *g*-factor, will also be high on both broad auditory perception and broad visual perception.

Disadvantages are as under:

(i) In factor analysis, there is occasional disagreement over the proper label for the factors. The technique itself only identifies factors and what they should be called is left to the researcher's intution and judgement.

(ii) Factor analysis can be good to the extent the obtained data allow. In psychology and education, where researchers often have to depend on less valid and reliable measures such as self-reports, ratings, etc., the technique can prove to be a problematic one.

(iii) In factor analysis, more than one interpretation can be made from the same data factored the same way. Factor analysis fails to identity this type of casuality.

☑ Review uestions

. What is normal curve Discuss the major characteristics of a normal curve.

2. Distinguish between parametric statistics and nonparametric statistics.

3. What is a Chi-square test Discuss its uses and applications.

. Distinguish between Type I error and Type II error in statistical significance. Also, e plain the role that level of significance plays in influencing the probability of making each of these two types of errors.

5. What is Yates' Correction Under what conditions is it applied

. Why is Student's *t* so named

. Discuss the difference between Correlation and Regression.

. Write short notes on the followings:
   (a) Correlation and regression
   (b) Rank difference correlation
   (c) Tuckey s SD
   (d) Multiple correlation

◆

24

WRITING A RESEARCH REPORT AND A RESEARCH PROPOSAL

CHAPTER PREVIEW

- General Purpose of Writing a Research Report
- Structure or Format of a Research Report
    Title Page
    Abstract
    Introduction
    Method
    Results
    Discussion
    References
    Appendix
    Author Note
- Style of Writing a Research Report
- Typing the Research Report
- Evaluating a Research Report
- Preparing a Research Proposal

GENERAL PURPOSE OF WRITING A RESEARCH REPORT

The writing of a research report is no less challenging a task than the research itself. It requires imagination, creativity and resourcefulness. Research reports should be written in a dignified and objective style, although there is no one such style which is acceptable to all.

The general purpose of a research report is not to *convince* the readers but to *let them know* what has been done, why it was done, what results were obtained and what the conclusions of the researcher were. Therefore, the research reports aim at telling the readers the problems investigated, the methods adopted, the results found and the conclusion reached. The research report should be written in a clear and unambiguous language so that the reader can also objectively judge the adequacy and validity of the research. For attaining objectivity, personal pronouns such as *I, you, we, my, our,* etc., should be avoided and as their substitutes, expressions like 'investigator', 'researcher' should be used. Needless to say, the highest standard of correct usage of words and sentences is expected.

To achieve these purposes is not an easy task. The matter is more complicated due to variations in style of writing research reports. However, the problem can be minimized if we adopt the style for research reports presented in the next section.

STRUCTURE OR FORMAT OF A RESEARCH REPORT (Style Manual)

The research report, whether it is based on a dissertation or short-term research paper, should be of fairly standardized pattern. There are different types of style manuals of reporting research. In fact, different departments or institutions develop their own style manual to which their theses or dissertations must conform. Careful examination, however, reveals that differences are concerned with minor details only. In general, all style manuals basically agree on the principles of correct and clear presentation. The style manuals commonly used in social sciences and humanities are: *The Publication Manual of American Psychological Association*; 5th ed. (2002); *The Chicago Manual of Style*, 16th ed. (2010); *the MLA Handbook for Writers of Research Papers*, 5th ed. (Gibaldi 1999); *Form and Style: Research Papers, Reports, Theses*, 11th ed. (Slade, 2000) as well as *A Manual for writers of Term papers, Theses and Dissertations*, 6th ed. (Turabian 1996). Of these various manuals, the most widely used style manual is the *Publication Manual*, 5th edition (2002) of American Psychological Association. The *Publication Manual* has given some excellent suggestions as to how to write a report effectively, how to present the basic ideas with an economy and smoothness of concise expression and how to avoid ambiguity and increase readaptability of the report. The following outline or format is the most popular typical research format prepared according to the American Psychological Association's *Publication Manual*, 5th ed. (2002).

I. Title Page
  A. Title
  B. Author's Name and Affiliation
  C. Running Head
II. Abstract
III. Introduction (no heading used)
  A. Statement of the Problem
  B. Background/Review of Literature
  C. Purpose And Rationale/Hypothesis
IV. Method
  A. Subjects
  B. Apparatus or instrumentation (if necessary)
  C. Design (if a complex design has been incorporated)
  D. Procedure
V. Results
  A. Tables and Figures
  B. Statistical Presentation
VI. Discussion
  A. Support Or Nonsupport of Hypotheses
  B. Practical and Theoretical Implications
  C. Conclusions
VII. References
VIII. Appendix (if appropriate)
IX. Author Note
  A. Identify each author's departmental affiliation
  B. Identify sources of financial support
  C. Acknowledge contributions to articles
  D. Give contact information for further information

Thus we find that the APA *Publication Manual* divides the format of writing a research report into nine parts. A discussion of these parts is given on the next page.

I. Title Page

The first page is the title page which basically contains three elements: title, author(s) and affiliation, as well as a running head. A discussion of these elements is given below.

A. Title and running head: The title should be concise and should clearly indicate the purpose of the study. It should not be stated so broadly that it seems to provide such an answer that can't be generalized either from the data gathered or from the methodology employed. Abbreviations should not be used in the title. The best length of a title is 12 to 15 words approximately. Introductory phrases such as *"A Study of..."* or *"An investigation of..."* should be avoided since it is understood that the investigator is studying something. For example, expression of the title like *"A Study of unlearning, spontaneous recovery and partial reinforcement effects in paired associate and serial learning"*, should be avoided.The preferred title would be *"Unlearning, spontaneous recovery and partial reinforcement effects in paired associate and serial learning."* The title should be typed in upper case and lower case letters centred on the page and when two lines are needed, they should be double-spaced.

A running head is an abbreviated or a short title which is printed at the top of the pages of a published article to identify the articles. On the title page the running head, that is, the short title which is a maximum of 50 characters including letters, punctuation and spaces between words, may be typed on the upper right side of the page. It is typed in all capital letters at the bottom of the title page. The running head in the above example may be *"Unlearning, Reinforcement and Learning."*

B. Author's name and affiliation: On the title page, the author's name should be centred below the title, and the next line should indicate the name of the institution to which the author is affiliated. In case of more than one author from the same institution, the affiliation should be listed last and only once.

UNLEARNING, REINFORCEMENT AND LEARNING
Unlearning, Spontaneous recovery, and Partial reinforcement effects
in Paired associate and Serial learning

ALOK K SINGH
PATNA UNIVERSITY, PATNA

I wish to acknowledge the assistance of the computer facilities rendered by the concerned institutions. I also wish to thank many students, teachers and principals of the various schools from where the reported data were collected. I also wish to thank K K Sinha, PhD, for his critical comments and B Sen, PhD, for his secretarial assistance.

Running head: UNLEARNING, REINFORCEMENT AND LEARNING

Fig. 23.1. An illustration of a title page

II. Abstract

The abstract is written on a separate sheet of paper which is page 2. It describes the study in about 100 to 150 words. In fact, the abstract is the summary of the study, which includes the problem under the study, method, such as characteristics of the subjects, research design, apparatuses, results (including statistical significance levels) and conclusions as well as implications. References must *not* be cited in the abstract.

III. Introduction

This starts on a fresh page and that is page 3. From this, the text of the report starts and all the sections of the test follow one another without a break, that is, Introduction, Method, Results and

Discussions follow one after another without any break and it is not necessary that they are started on new pages.

Page 3, where Introduction starts, is not labelled as 'Introduction'. However, the running head on the upper right side of the page and the complete title without the name of the author in the centre of the page are entered. A good introduction has three components.

(i) The researcher must give a clear and definitive statement of the problem and must develop it logically in the light of pertinent studies. He must indicate the need for the present research and why the problem is important in terms of theory and/or practice.

(ii) The second important component of introduction is the review of the previous literature. The researcher must try to establish an understanding of the existing literature relevant to his or her study. He needs to connect logically the previous body of literature with the present study.

(iii) The third and final component of the introduction is to formulate a clean rationale of the hypotheses to be proposed. Every hypothesis must be clearly stated so that it is clear how it would be scientifically tested. Different variables and terms should be properly defined and investigated.

IV. Method

The main body of the report continues with the method section after introduction. The major purpose of this section is to tell the reader *how* the research was conducted. The method section is separated from introduction by the centred `Method'. The subsections are subsequently labelled at the left margin and underlined. In general, the following subsections, in order, are included in the method section.

A. *Subject*: The population should be clearly specified as well as the method of drawing samples should also be stated. In specifying the population, such characteristics like age, sex, geographical location, SES, race, institutional affiliation, etc., should be clearly mentioned. Not only that, the total number of subjects and the number assigned to each condition should also be spelled out. The researcher should also state that the treatment given to subjects was in accordance with the ethical standards of APA.

B. *Apparatus*: If the research has been conducted with the help of some relevant apparatuses, their names and model numbers should also be mentioned so that another researcher, if he wishes, may obtain or construct it. If possible, the researcher should provide a diagram of the apparatus in his write-up and this is extremely important where the apparatus is a complex and novel one. Minor pieces of apparatus such as pencil, pen, blade, etc., should not be listed.

C. *Design*: The type of research design should also be spelled out after the subsection of apparatus.* Here the procedure for assigning participants to groups as well as labels to groups (experimental or control) should also be indicated. IV and DV should be clearly spelled out and these variables should be carefully defined if they have not been defined in the introduction section. The technique of exercising experimental control should also be spelled out.

D. *Procedure*: This subsection describes the actual steps carried out in conducting the study. This includes the instruction given to the subjects (if they are human), how IV was administered and DV was measured. It also includes assignment of subjects to conditions and order of assessments if more than one. Anyway, enough information must be provided to permit easy replication by the subsequent researcher.

---

* If the research does not incorporate a complex design, the 'design' subsection may be dropped and rubrics of design may be included in Procedure.

V. Results

The third section of the main body is the results whose purpose is to provide sufficient information about how the conclusion was reached. The heart of this section is the presentation of data relevant to test the hypothesis. All relevant data are presented including those that don't support the hypothesis. *Tables* and *Figures* are commonly employed for supplementing textual material. A table consists of several numbers that summarize the major findings of the experiment. A figure is a graph, photograph, chart or like materials, which are relevant particularly for certain kinds of data like showing the progress of learning or maturation over a designated period. Data in the text and in tables or figures should not be redundant, rather, they should be complementary. Results of the statistical analyses carried should be provided and the level of significance for these statistical analyses should also be presented. However, they should not be interpreted and discussed in this section.

VI. Discussion

The final section of the main body report is discussion. The major function of this section is to *interpret* the results of the study and to *relate* those results to other studies. The implications of the study including the hypotheses, supported or not supported, are discussed. If the findings are contrary to the hypothesis, some new explanation is required so that a new hypothesis may be advanced. New hypotheses may also be advanced about any uncommon deviation in the results. Sometimes faulty hypotheses can be modified to make them consistent with the results obtained. Negative results should also be discussed. Such results occur when a hypothesis predicts something but the results don't support that prediction. A brief speculation about why they occurred is sufficient. A brief discussion of the limitations of the present study and proposals for future research is appropriately discussed here. Here the researcher finally includes conclusions that reflect whether the original problem is better resolved as a result of the investigation.

VII. References

The `References' section begins at a new page with the label 'References' at the centre. References comprise all documents including journals, books, technical reports, computer programms and unpublished works mentioned in the text of the report. References are arranged in alphabetical order by the last name of the author(s) and the year of publication in parenthesis or in case of unpublished citations, only the reference is cited. Sometimes no author is listed and then, in that condition the first word of the title or sponsoring organization is used to begin the entry. When more than one name is cited within parenthesis, the references are separated by semicolons. In parenthesis, the page number is given only for direct quotations. The researcher should check carefully that all references cited in the text appear in the references and vice versa.

References should not be confused with *Bibliography*. A bibliography contains everythin that is included in the reference section *plus* other publications which are useful but were no cited in the text or manuscript. Bibliography is not generally included in research reports. Onl references are usually included. The *Publication Manual* of the American Psychologic Association, 5th ed. (2002) has given some specific guidelines for writing references of variou types of works as indicated below.

1. For references of books with single author:
   Siegel, S (1956) *Nonparametric Statistics for the Behavioural Sciences*, New Yo McGraw-Hill.
2. For references of books with multiple authors:
   Guilford, J P & Fruchter, B (1978) *Fundamental Statistics in Psychology And Educati* New York: McGraw-Hill.
3. For references of Editor as author:
   Misra, G. (ed.) (2011), *Psychology in India*. vol., 4. Delhi: Pearson.

4. For references of corporate or association as author:
   American Psychological Association (1983) *Publication Manual* (3rd ed.) Washington.
5. For references of journal article:
   Edwards, A L & Kenny, KC (1946). A comparison of Thurstone and Likert techniques of Attitude Scale Construction, *Journal of Applied Psychology, 30*, 72–73.
6. For references of thesis or dissertation (unpublished):
   Singh, AK (1978) *Construction And Standardization of a Verbal Intelligence Scale.* Unpublished Doctoral Dissertation, Patna University, Patna.
7. For reference of chapter in an edited book:
   Atkinson, R C & Shiffrin, R M (1968) Human Memory: A proposed system and its control processes. In *The Psychology of Learning and Motivation*, ed. K W Spence and J T Spence, vol. 2, pp 89–195. New York: Academic Press.
8. For references of magazine articles: Basant., K. (2013, August). The usefulness of psychology to common people. *Psychology Today and Tomorrow.* 15–20.
9. For references of Unpublished paper presented at a meeting:
   Singh, A. K. Kumar, P & Sen Gupta, A (2005, July) *Demographic Correlates of Job stress.* Paper presented at the Annual Meeting of Council for Social Sciences, Lucknow.
10. For references of unpublished manscripts:
    Gupta, R. (1995) Social awareness in relation to media among college students. unpublished Manuscript, Ranchi University, Ranchi.

    Orne, M. K. (2013) *A scale to assess Big Five Personality dimensions.* Manuscript submitted for Publication.
11. References of citations from internet sources:
    American Psychological Association (2001) APA style Homepage. Retrieved, July 30, 2001, from http://www.apastyle.org/.

VIII. Appendix

In an *appendix,* those items of information are provided that would be inappropriate or too long for the main body of research. Each appendix appears as a new page with the label 'Appendix' along with the identifying letter, centred. Usually in an appendix, materials such as tests, detailed statistical treatments, computer programme, etc., are included.

IX. Author Note

In this part the researcher writes about four basic things including source of financial support, acknowledgement, contact information and each author's affiliation, in case many authors have contributed to the research work.

STYLE OF WRITING A RESEARCH REPORT

The research report should be written in a style that is clear and concise. Slang, hackneyed and folksy style should be avoided. The research report should describe and explain rather than try to convince or move to some action. Personal pronouns should be used when they are most appropriate and demanding; otherwise their use should be discouraged to ensure proper objectivity. Only the last name(s) of the cited author(s) should be used. Titles like Dr, Mr, Professor, Dean, etc., should be omitted. In describing research procedures, the past tense should be used. Abbreviations should be used only after their referent has been clearly spelled out in parenthesis. Of course, there are a few exceptions to this rule for well-known abbreviations such as IQ or DIQ.

Important and standard statistical formulas are not presented in the research report. Detailed statistical computations are also not included. However, if some unusual formulas are used in statistical analysis, it is appropriate to mention them.

Numbers beginning a sentence should always be spelled out. Fractions and numbers of less than ten should be spelled out. The expression `one-half' is preferred but for all fractions the expression should be like "6.5" or "61/2". The word `per cent' (which means per hundred) should be spelled out unless it is in tables and figures. Arabic numerals with per cent such as 20 per cent should be used unless they begin a sentence. Commas should clearly indicate thousands or millions such as 20,305,682.

Apart from these specifications, some general rules for writing a research report/article that conforms to APA style should be followed. These general rules may be discussed as under:

1. Since a research report describes a completed study, it should be written in the past tense. The exception is to write in the present tense only the conclusions that apply to present or future situations.

2. All sources from which information are obtained are cited by only the last names of the authors and the year. Reference can be used as the subject of the sentence: Hilgard and Bower (1975) defined learning "...". Or the researcher may state an idea and provide the reference in parentheses: Learning is defined as ... (Hilgard & Bower 1975). In parenthetical citation '&' is used in pace of "and". When there are three to six authors, all names should be included at the first time of citation and subsequently, it can be referred to by using the first author and the Latin phrase *et al.* Thus, in the first citation, we can write Hilgard, Bower and Atkinson (1999), but subsequently, we can say "Hilgard *et al.* (1999)". If there are more than six authors, we can cite the name of the first author and *et al.* even the first time.

3. Abbreviations should be avoided. They are justified only in (a) it appears very frequently throughout the report, (b) a term consists of several words, (c) numerous different abbreviations are not being used. If an abbreviation has to be used, this should be done by creating an *acronym,* using the first letter of each word of the term. The complete term should be defined the first time it is used, with its acronym in parentheses. For example, Long-Term Memory (LTM) is defined as "...". Subsequently only the acronym, except as the first word of a sentence, should be used. As a first word of a sentence, the complete term should be used.

4. As far as possible, direct quoting should be avoided. The idea should be paraphrased and summarized so that readers should know what the authors intend to say. Besides, attempt should be made to address a study itself, not its authors. For example, the phrase "Hilgard et al." mainly refers to a reported study and not to the people who conducted it. In this way, we should write "the results are reported *in* Hilgard *et al.* (1999) instead of "The results are reported *by* Hilgard *et al.* (1975)".

5. For distinguishing your study from other studies, refer to as "this study" or "the present study".

6. As far as possible, approved psychological terminology should be used. When a nonstandard term or name of a variable is to be used, define the term or word for the first time and then use that word consistently.

7. Numbers between zero and nine should be written in words and numbers that are 10 and larger should be written in digits. However, digits can be used for any size number if (a) the author is writing a series of numbers in which at least one is 10 or larger, or (b) the number expresses a decimal or refers to a statistical result or to a precise measurement. *No sentence should begin with a number expressed in digits.*

8. Although in most research reports the generic term *subjects* are used, the APA style prefers the use of a less intentional and more precise term for it. Therefore, *Participants* rather than subjects should be written but where appropriate, more descriptive terms such as *men, women, children, pigeons,* etc., should be used.

If a research report is written keeping the above general rules in view, it would definitely provide readers the information necessary for (a) understanding the study, (b) evaluating the study and (c) performing a literal replication of the study.

TYPING THE RESEARCH REPORT

It is the responsibility of a research writer to present manuscript material to the typist in an ordered and proper form. Except for minor typographical matters, the correction of major errors is not the responsibility of the professional typist. However, while typing a research report, the following guidelines should carefully be followed:

(1) A good quality paper or bond paper, preferably of $8\frac{1}{2}$" and 11" size, should be used. Only one side of the sheet is used for typing the materials.

(2) All margins, that is, top, bottom, left and right should be $1\frac{1}{2}$ inches.

(3) Words should not be split at the end of the line unless completing them would definitely interfere with the margin. If a word is to be divided at all, a dictionary for correct syllables should be consulted.

(4) Direct quotation of three or less than three typewritten lines are included in the text and enclosed in quotation marks. Quotations of larger length (that is, more than three typewritten lines) are set off from the text in a double-spaced paragraph and indented about five spaces from the left margin without quotation marks.

(5) Pagination information is given in parenthesis at the end of the direct quotation.

(6) Words, letters, digits or sentences which are to be finally printed in italics should be underlined.

EVALUATING A RESEARCH REPORT

A critical analysis of a research report is a valuable aid for understanding and gaining insight into the nature of problems, methods adopted for solution of the problems, the ways in which data are processed and conclusions reached. For making a critical evaluation of a research report, the following questions may be raised.

1. *The title:* (a) Is it clear? (b) Is it concise?

2. *The problem:* (a) Is it clearly stated? (b) Is its significance recognized? (c) Have specific questions been raised? (d) Are testable hypotheses framed? (e) Are assumptions stated? (f) Are limitations recognized? (g) Are important terms defined?

3. *Review of the literature:* (a) Does it well cover the area? (b) Are the main findings of the previous researches noted? (c) Is it well organized and summarized?

4. *Methods:* (a) Is the research design appropriate? (b) Is the research design described in detail? (c) Are the samples adequate? (d) Are the extraneous variables well organized and controlled? (e) Are the psychometric properties of data-gathering instruments or questionnaires satisfactory? (f) Are the statistical tests appropriate?

5. *Results and discussion:* (a) Are the appropriate uses of tables and figures made? (b) Is the discussion in the text clear and concise? (c) Is the data analysis logical and perceptive? (d) Is the statistical analysis correctly interpreted?

6. *Summary and conclusion*: (a) Is the problem reinstated? (b) Are the questions or hypotheses restated? (c) Are the procedures well described? (d) Are the findings concisely reported? (e) Are the findings and conclusions based upon the data collected and analyzed?

Using the pattern suggested above, the reader can make a critical analysis of any research report which would, in turn, help him in developing competency in conducting a research and preparing a good report.

PREPARING A RESEARCH PROPOSAL

The writing of a research proposal is an important aspect of the research process. Before we elucidate the steps involved in preparing a research proposal, it is essential to throw light upon the nature and need for a research proposal.

A research proposal is a detailed plan of the research to be conducted. It comparable to the blueprint which the architect prepares before the work of building commences. The research proposal provides the basis for evaluation of the submitted project. Many institutions require that a proposal must be submitted before it is finally approved. Researchers often seek financial support for their research work by submitting grant proposals to the government and even private agencies. The major purpose of a research proposal is to ensure a workable experimental design which, when implemented, may result into analyzable and interpretable piece of research or finding of significant scientific merit.

Generally, a written research proposal follows the general format of a journal article but headings of various sections are a bit different. The following are the nine steps that are generally followed in preparing a research proposal. (Many institutions or agencies may suggest some other formats for the research proposals.)

1. Problem
2. Definitions, assumptions, limitations or delimitations
3. Review of related literature
4. Hypothesis
5. Methods
6. Time schedule
7. Expected results
8. References
9. Appendix

These nine steps may be discussed as follows:

1. *Problem:* The problem of the research proposal is expressed in a declarative statement but it may be in the question form. The problem must be stated in such a form that it clearly tells about the major goal of the research. Besides its formulation, the researcher must mention why it is worth the time, effort and expense required to conduct the proposed research. In other words, the proposal writer should not only mention the problem clearly but must also demonstrate its significance. A few examples of problem statements are as follows:

(a) Coeducation improves the morality level of students.

(b) Active participation by students in politics may have damaging effect upon their creativity.

2. *Definitions, assumptions, limitations and delimitations*: The proposal writer must define all the important variables included in the study in operational terms. These definitions provide a good background with which the researcher approaches the problem. Assumptions implied in the proposed study should also be clearly mentioned. Assumptions are statements which the researcher believes to be a fact but he can't verify them. Limitations of the study should also be clearly mentioned. Limitations generally include those factors or conditions which are beyond the control of the researcher but are likely to influence the conclusions or findings of the study. Inability to randomly select subjects for the study and unavailability of reliable and valid data-gathering instruments can be good limitations of any research. Delimitations of the proposed research should also be clearly spelled out. Delimitations refer to the boundaries of the study. In other words, delimitations clearly tell who will be included as the sample of the study and for whom the obtained conclusion will be valid.

3. *Review of related literature*: The research proposal should include a more extensive review of the relevant literature. An effective relevant literature includes those studies which have been competently executed and clearly reported and are closely related to the present

problem. This step ensures that the researcher is familiar with what is already known and what is still unknown and has to be verified and tested. Moreover, it also helps to eliminate the duplication of what has already been done and provides the background for useful suggestions for further investigations. In search of related literature, the researcher, among others, should concentrate upon similar but competently executed studies, design of the study, sampling methods, population sample, variable defined, extraneous variables controlled, recommendations for further research, etc.

4. *Hypotheses*: The research proposal should include the major hypotheses to be tested. Some minor hypotheses, if any, should also be formulated. Since a research hypothesis is a tentative answer to a question, it is important that the hypothesis should be formulated before data are gathered. In fact, the formulation of hypothesis is such a step which clarifies the nature of the problem and also the underlying logic of the research investigation. The hypothesis should be reasonable, consistent with known facts in the concerned area, testable and such that it can be stated in the simplest possible terms.

5. *Methods*: This part of the research proposal is very important. It includes three subsections—subjects, procedures and data analysis. The subjects subsection spells out the details of the population from which subjects are to be selected. The total number of subjects desired from the population and how they will be selected are generally indicated in this subsection. The procedure subsection outlines the details of the research plan. In other words, this subsection outlines in detail what will be done, how it will be done, what data will be needed and what data-gathering devices will satisfactorily be used. The data-analysis subsection outlines the details of the method of analyzing data by different statistical techniques. The details should preferably mention the rationale behind selecting the statistical techniques.

6. *Time schedule*: An effective research proposal must have a clear time schedule in which the entire subject should be divided into manageable parts and probable dates should be assigned for their completion. Such steps help the investigator in budgeting his time and energy effectively, systematizing the study and minimizing the tendency to delay the completion.

7. *Expected results*: A good research proposal should also indicate the possible or expected results as far as possible, although in some cases it may prove to be a Herculean task for the investigator to spell out the expected results. The expected results section should include a brief discussion of the anticipated results of the research and should also highlight those that are the most important for the research. In this section reasonable alternative to the expected results should also be mentioned as well as those likely problems should also be spelled out which may originate if the results show deviation from the research hypothesis.

8. *References*: The reference section should include the names of the authors along with the details of the publication of their research work. It should be more or less like that same section as it would be submitted with the final report. Sometimes it is just possible that the literature may have to be included in the `Discussion' section of the final report that was not anticipated in the proposal. But this should be an exception and not the rule.

9. *Appendix*: A research proposal ends with an appendix. An appendix should include a list of all materials that are to be used in the study. Among other things, it may include a copy of the test or scale used, list of stimulus materials and apparatuses, a copy of instructions to be given to the subjects, and so on.

Thus we find that the research proposal has several steps before it reaches its completion. A good research proposer must keep all these steps in view at the time of writing a research proposal.

☑ Review Questions

1. Taking an example of any research project, outline the major steps in writing a research report.
2. Discuss fully the main points to be considered in outlining a research proposal.
3. Discuss some of the important considerations to be kept in mind while writing a research report.

◆

Objective Questions

1. Introduction to Measurement

1. Which is the correct meaning of measurement?
   (a) Measurement means to measure something correctly.
   (b) Measurement means assigning of numerals.
   (c) Measurement means assigning numerals according to some rules.
   (d) Measurement means assigning numerals arbitrarily.
2. Which is NOT a property of measurement?
   (a) Measurement involves assigning numerals according to rules.
   (b) Measurement is concerned with some attributes or features.
   (c) In measurement, numerals are used to represent quantities of attribute.
   (d) In measurement, numerals are used to represent qualities of objects.
3. Who invented the term mental test?
   (a) Galton
   (b) Cattell
   (c) Goodenough
   (d) Pearson
4. If $a = p$ and $b = q$ then $a + b = p + q$. This postulate of measurement tells that
   (a) the relationship between a and b is asymmetrical.
   (b) in the process of adding, identical numbers may be substituted for each other without making change in the result.
   (c) in the process of addition, the order of combinations of objects or numbers makes no difference.
   (d) the objects which are equal to the same object are also equal to one another.
5. Which is the highest level of measurement?
   (a) Nominal measurement
   (b) Ordinal measurement
   (c) Interval measurement
   (d) Ratio measurement
6. In which of the following scales are all the three basic properties such as magnitude, equal interval and absolute ratio incorporated?
   (a) Nominal scale
   (b) Ordinal scale
   (c) Interval scale
   (d) Ratio scale
7. Which can't be incorporated as one of the functions of measurement?
   (a) Selection function
   (b) Classification function
   (c) Review function
   (d) Comparison function
8. Which is not a true point of distinction between psychological measurement and physical measurement?
   (a) In physical measurement, the unit of measurement is not fixed and constant but in psychological measurement this is fixed and constant.
   (b) In physical measurement there is true zero point, whereas in psychological measurement there is an arbitrary zero point.
   (c) Physical measurement is more accurate and predictable than psychological measurement.
   (d) Physical measurement is direct whereas psychological measurement is indirect.

Scanned with CamScanner

9. Which is not a correct source of error in measurement?
   (a) Respondent (b) Measurer
   (c) Test situation (d) Presence of people
10. Which is not a general problem of measurement?
   (a) Relativity of measurement (b) Incompleteness of measurement
   (c) Indirectness of measurement (d) Randomness in measurement

Answers

1. (c) 2. (d) 3. (b) 4. (b) 5. (d) 6. (d) 7. (c)
8. (a) 9. (d) 10. (d)

2. Test Construction

1. Which is a correct meaning of psychological test?
   (a) Psychological test measures traits/abilities.
   (b) Psychological test is a standardized measure of sample of behaviour.
   (c) Psychological test is only a qualitative measure of behaviour.
   (d) Psychological test is only a quantitative measure of behaviour.
2. Which is not a characteristic of good psychological test?
   (a) Objectivity (b) Reliability
   (c) Norms (d) Comprehensibility
3. By standardization of test is meant
   (a) there must be a standard way of giving instructions for maintaining uniformity
   (b) test must have reliability, validity and norms
   (c) test must have uniformity in scoring
   (d) all of the above
4. A test in which there is a generous time limit so that most examinees are able to attmept it, is known as
   (a) verbal test (b) speed test
   (c) power test (d) nonverbal test
5. The major feature of nonverbal test is that
   (a) it contains pictorial items
   (b) it contains only manipulative items
   (c) it contains both pictorial as well as manipulative items
   (d) it contains only verbal items
6. Which is not a characteristic of good psychological and educational test?
   (a) Such test is based upon unlimited sample of behaviour.
   (b) Such test usually provides scores or categories, which are subsequently interpreted.
   (c) Such test provides both quantitative and qualitative measurement.
   (d) Such tests are either norm-referenced or criterion-referenced.
7. The problem part of the item in objective test is known as
   (a) stem (b) response option
   (c) response bias (d) question
8. The other name of free-answer test is
   (a) objective test (b) essay test
   (c) speed test (d) power test
9. Which is not a correct use of psychological test?
   (a) In classification (b) In determining ethical issues
   (c) In diagnosis (d) In self-knowledge

Scanned with CamScanner

10. Which can't be considered as one of the limitations of a psychological test?
    (a) Psychological test represents an invasion of privacy.
    (b) Psychological test penalizes bright and creative examinees.
    (c) Psychological test creates anxiety.
    (d) Psychological test fails to establish a close contact with examinees.

Answers

1. (b) 2. (d) 3. (d) 4. (c) 5. (a) 6. (a) 7. (a)
8. (b) 9. (b) 10. (d)

3. Item Writing

1. Who has said that items are the lowest common denominator of a test?
   (a) Bean (b) Anastasi
   (c) Nunnally (d) None of these
2. For assessing noninterpretative information, what type of item is most suited?
   (a) Multiple-choice item (b) Two-alternative time
   (c) Matching item (d) None of these
3. For assessing attitude of one person, which type of item formats are frequently used?
   (a) Checklist (b) Q-sort
   (c) Likert format (d) Category format
4. Identification items are also known as
   (a) supply-type item (b) selection-type item
   (c) completion-type item (d) none of these
5. Which can't be included as one of the best characteristic of item of the test?
   (a) Item should not contain ambiguity.
   (b) Item should be easy.
   (c) Item should not present any difficulty in reading.
   (d) Item should have discriminating power.
6. Which one can't be included in the limitations of the matching item?
   (a) Matching items do not measure real understanding of concepts.
   (b) Matching items provide opportunity to arrive at the correct answer through the process of elimination.
   (c) Matching items are not suited when the purpose is to measure different types of information.
   (d) Matching items require longer vision and depth of knowledge in arriving at the correct answer.
7. What is the major purpose of including negative and positive items in the test?
   (a) To maintain a balance
   (b) To promote practicality in the test
   (c) To check acquiescence response set
   (d) To enhance reliability of the item
8. For scoring objective items, which method is the most appropriate and correct?
   (a) Overlay-key method (b) Likert-key method
   (c) Marshall's item response method (d) Anastasi-key method
9. Double-barreled items are those item, which convey
   (a) two or more ideas at the same time
   (b) two or more attitudes at the same time

Scanned with CamScanner

(c) two or more emotions at the same time
(d) two or more feelings at the same time

10. An item of the test is best defined as a single task that
(a) can be partially broken down into smaller units
(b) can be wholly broken down into smaller units
(c) can't be broken down into any smaller units
(d) can be stretched and broken down into two smaller units

Answers

1. (c) 2. (b) 3. (c) 4. (b) 5. (b) 6. (d) 7. (c)
8. (a) 9. (a) 10. (c)

4. Item Analysis

1. Which is not the correct purpose of item analysis?
(a) It indicates effectiveness of distractors.
(b) It provides information about index of difficulty only.
(c) It provides opportunity to make a particular item more functional.
(d) It provides information about index of discrimination.

2. The proportion giving correct answers to an item is called as
(a) difficulty value of item
(b) discrimination value of item
(c) correct distractor
(d) none of these

3. When the proportion of passing an item in the upper group is lower than the proportion of passing the same item in the lower group, the test constructor tries to
(a) bring modificaiton in the item
(b) reject such item from inclusion
(c) accept such item for final inclusion
(d) do none of these

4. Which is not an appropriate statistical technique for determining the discrimination index of the item?
(a) Product-moment correlation
(b) Biserial *r*
(c) *Phi*-coefficient
(d) Kendall's *tau*

5. The difficulty of an item is influenced by several factors. Which one can't be included in this list?
(a) Ambiguity of the item
(b) Past experience of examinee
(c) Homogeneity among alternative responses
(d) Dichotomy in item

6. The steepness or slope of the item characteristic curve gives information about
(a) difficulty value of the item
(b) discriminating power of the item
(c) effectiveness of distractors of item
(d) none of these

7. What is the other name of item response theory?
(a) Latent trait theory
(b) Latent-type theory
(c) Latent content theory
(d) None of these

8. When the item-total correlation is negative, the slope of item characteristic curve is
(a) positive
(b) negative
(c) flat
(d) not known

9. When the difficulty value of an item is 1.00 or 0 (zero), the discriminative value of the item will be
(a) high
(b) low
(c) negligible
(d) none of these

Scanned with CamScanner

10. Which is not a correct statement?
    - (a) Distractor affects discriminative ability of the item.
    - (b) Distractor affects difficulty value of the item.
    - (c) Distractor affects neither discriminative ability nor difficulty value of the item.
    - (d) Distractor creates complexity in the meaningfulness of the item.

Answers

1. (b) 2. (a) 3. (b) 4. (d) 5. (d) 6. (b) 7. (a)
8. (b) 9. (c) 10. (c)

5. Reliability

1. Reliability is defined as
    - (a) self-correlation of the test
    - (b) correlation of the test with some criteria
    - (c) proportion of error variance in total variance
    - (d) none of these
2. Which one is true in case of a highly reliable test?
    - (a) Such test yields an inconsistent result.
    - (b) Such test yields a consistent result.
    - (c) Such test produces higher error variance.
    - (d) None of the above
3. Logically, reliability is defined as proportion of
    - (a) true variance in total variance
    - (b) error variance in total variance
    - (c) zero variance in total variance
    - (d) high error and true variance in total variance
4. Which does not assess the internal constistency reliability?
    - (a) Test-retest reliability
    - (b) Split-half reliability
    - (c) K–R formulae
    - (d) Cronbach's alpha
5. When you have developed a test containing a multiple scored items and you are interested in calculating internal consistency reliability, which method will you adopt?
    - (a) Split-half method
    - (b) Cronbach's alpha
    - (c) K–R formula
    - (d) Rulon formula
6. When reliability of half test is 0.70, the reliability of test after Spearman-Brown prophecy formula would be
    - (a) 0.90
    - (b) 0.75
    - (c) 0.82
    - (d) 0.68
7. If the difficulty values of all items of a test is the same, the most appropriate technique of calculating reliability of test would be
    - (a) K–R 20
    - (b) K–R 21
    - (c) Rulon formula
    - (d) Flanagan formula
8. When the reliability coefficient of the test is 0.65, its index of reliability will be
    - (a) 0.76
    - (b) 0.81
    - (c) 0.92
    - (d) 0.93
9. If you are calculating reliability of difference score, then higher correlation between X and Y will produce
    - (a) higher reliability of difference score
    - (b) lower reliability of difference score

Scanned with CamScanner

(c) moderate reliablity of difference score
(d) none of these

10. Which does not affect the reliability of the test?
(a) Difficulty value of the item
(b) Length of the test
(c) Homogeneity of the test
(d) Standard error of measurement

11. When the reliability coefficient of a test is 1.00, the standard error of measurement will be
(a) 1.00
(b) 0 (zero)
(c) 0.50
(d) none of these

12. What is the source of error variance in split-half reliability?
(a) Item sampling
(b) Time sampling
(c) Double sampling
(d) None of these

13. Which one is equivalent of the average of all possible split-half reliability coefficients of the test?
(a) Cronbach's alpha
(b) Rulon formula
(c) K–R formula
(d) None of these

Answers

1. (a) 2. (b) 3. (a) 4. (a) 5. (b) 6. (c) 7. (b)
8. (b) 9. (b) 10. (d) 11. (b) 12. (a) 13. (a)

6. Validity

1. Which is not the correct property of validity of the test?
(a) Validity is a relative term and therefore, a test is not generally valid.
(b) Validity is not a fixed property of the test.
(c) Validity is a matter of degree.
(d) Validity is the self-correlation of the test.

2. A testing enterpreneur has developed a Big Toe Intelligence test. For assessing your intelligence, he measures your right big toe in centimeters, multiplies that figure by 100 and adds your age. Then he tells about your IQ. Given your knowledge of testing, you know this test is
(a) reliable but not valid
(b) not reliable
(c) valid
(d) probably standardized for abnormal people and not for normal ones

3. If any psychological or educational test yields the same results consistency but does not measure what it intends to measure, it is
(a) valid but not reliable
(b) reliable but not valid
(c) neither reliable nor valid
(d) both reliable and valid

4. An achievement test must have
(a) concurrent validity
(b) content validity
(c) construct validity
(d) none of these

5. A content validity requires that
(a) the test items must have sampling validity only
(b) the test items must have item validity only
(c) both 'a' and 'b' should be there
(d) neither 'a' nor 'b' is needed at all

Scanned with CamScanner

6. If a newly constructed intelligence test is being correlated with an already standardized test of intelligence and if it yields high correlation coefficient, it becomes evidence for
    (a) predictive validity
    (b) construct validity
    (c) concurrent validity
    (d) factorial validity
7. Which of the following does not affect validity of the test?
    (a) Length of the test
    (b) Sample heterogeneity
    (c) Ambiguous direction
    (d) Standard error
8. If the reliability of the test is 0.80 and the reliability of the criterion is 1.0, the maximum validity would be
    (a) 0.75
    (b) 0.82
    (c) 1.23
    (d) 0.89
9. When validation of a completed test is done on a new sample taken from the same population, the process is known as
    (a) cross-validation
    (b) cross standardization
    (c) cross-hetrogeneity
    (d) cross divergence
10. When a test correlates with its expected referents, the process is known as ...... and when a test correlates poorly with measures with which it should not, the process is known as ....
    (a) convergent validity, construct validity
    (b) divergent validity, convergent validity
    (c) divergent validity, content validity
    (d) convergent validity, divergent validity

Answers

1. (d) 2. (a) 3. (b) 4. (b) 5. (c) 6. (c) 7. (d)
8. (d) 9. (a) 10. (d)

7. Norms and Test Scales

1. By formulation of ......, we are able to standardize the use and standardization of a given test.
    (a) validity
    (b) reliability
    (c) norms
    (d) temporal stability
2. When a test score is interpreted by establishing an external standard, such test is known as
    (a) norm-referenced test
    (b) criterion-referenced test
    (c) construct validity
    (d) none of these
3. Suppose you are going to construct test norms for general intelligence of children belonging to the age range of 3 years to 16 years. In such a situation, the most appropriate norms would be
    (a) grade norms
    (b) age norms
    (c) percentile norms
    (d) none of these
4. Deviation IQ is an example of
    (a) T score
    (b) z score
    (c) stanine score
    (d) standard score
5. If the obtained score is 80, standard deviation is 10 and mean is 60, then z score will be
    (a) +3.0
    (b) +2.0
    (c) +1.75
    (d) +2.50

Scanned with CamScanner

6. Which is not a limitation of percentile norms?
   (a) Percentile and percentage score are confused.
   (b) In percentile norms, there remains inequality of units throughout the percentile scale.
   (c) Percentile norms indicate only the person's relative position in the standardization sample.
   (d) Percentile norms do require a big sample.
7. In stanine score the mean is ...... and the standard deviation is ....
   (a) 5, 1.96 (b) 6, 2
   (c) 5, 2.96 (d) 4, 2.96
8. Which of the following statements is correct?
   (a) Deviation IQ is an improvement over traditional IQ.
   (b) Deviation IQ is calculated by MA/CA × SD.
   (c) Deviation IQ is a standard score.
   (d) The concept of deviation IQ was proposed by Stern.
9. When transformation of raw score in done in such a way that all the characteristics of original distribution of raw scores are retained without any change, this process is known as
   (a) linear transformation (b) normalized standard score
   (c) cumulative frequency distribution (d) none of these
10. The standard about standard score is that
    (a) it has a fixed mean and floating standard deviation
    (b) it has a fixed mean and fixed standard deviation
    (c) it has fixed standard deviation but floating mean
    (d) it has both floating mean and floating standard deviation.

Answers

1. (c)	2. (a)	3. (b)	4. (d)	5. (b)	6. (d)	7. (a)
8. (c)	9. (a)	10. (b)				

8. Response Set in Test Scores

1. The deviation set was proposed by
   (a) Berg (b) Cronbach
   (c) Anastasi (d) Freeman
2. When the testees or examinees give a very unusual or uncommon response, it is known as
   (a) cautiousness (b) deviation set
   (c) evasiveness (d) extremeness
3. In the popular Rorschach test, which of the following response sets is commonly observed?
   (a) Faking-good (b) Faking-bad
   (c) Deviation set (d) Evasiveness
4. When an examinee deliberately paints a negative picture of the self on any test, this tendency is known as
   (a) social desirability tendency (b) social undesirability tendency
   (c) social altruistic tendency (d) none of these
5. When the difficulty value of the items of a test is low such as 0.10 or 0.15, the examinee is most likely to exhibit
   (a) tendency to work speedly (b) semantics
   (c) tendency to guess (d) coutiousness

Scanned with CamScanner

6. When the testee is found to respond in some systematic way by choosing any one type of the response options in each item, he is definitely showing
   (a) deviation set
   (b) acquiescence set
   (c) faking-good
   (d) faking-bad

7. Which statement is not true?
   (a) Response sets widen the range of individual differences.
   (b) Response sets contribute to the error variance of the test.
   (c) Response sets are usually seen in those tests whose items are vague and unstructured.
   (d) Response sets show consistency.

8. When two-alternative response option items are written, the test is likely to suffer from
   (a) acquiescence, evasiveness
   (b) evasiveness, extremeness
   (c) faking-good and acquiescence
   (d) evasiveness and faking-bad

9. For controlling acquiescence fully, psychologists recommend
   (a) two-point scales of judgement
   (b) three-point scales of judgement
   (c) five-point to seven-point scales of judgement
   (d) none of these

10. Researchers have revealed that social disability tendency is related to
   (a) need for social approval
   (b) need for social conforming
   (c) need for avoiding social criticism
   (d) all of these

Answers

1. (a) 2. (b) 3. (c) 4. (b) 5. (c) 6. (b) 7. (a)
8. (a) 9. (c) 10. (d)

9. Measurement of Intelligence, Aptitude and Achievement

1. Who said, "Intelligence is whatever intelligence test measures".
   (a) Thurstone
   (b) Binet
   (c) Boring
   (d) Spearman

2. Which of the following has not been included in Thurstone's primary mental ability test?
   (a) Word fluency factor
   (b) Reasoning factor
   (c) Rote memory factor
   (d) Semantic factor

3. A positive correlation between two or more than two functions indicates the presence of
   (a) *g*-factor
   (b) *s*-factor
   (c) *r*-factor
   (d) *z*-factor

4. Piaget has defined intelligence on the basis of
   (a) assimilation and approximation
   (b) assimilation and accommodation
   (c) accommodation and approximation
   (d) approximation and approval

5. An expert in the field of intelligence says that there are three different kinds of intelligence, distinguished by three factors: the content of one's knowledge, the product that represents it and the operation that the person performs. This expert agrees with the viewpoints of
   (a) Spearman
   (b) Guilford
   (c) Goddard
   (d) Gardner

6. If a 10 years old has a mental age of 14 years, his IQ would be
   (a) 140
   (b) 130
   (c) 120
   (d) 100

Scanned with CamScanner

7. The largest percentage of people taking an IQ test will score in range from
   (a) 100 to 140 (b) 80 to 100
   (c) 90 to 109 (d) 60 to 90
8. Who introduced the fifth edition of Stanford–Binet test (SB-5)?
   (a) Thorndike (b) Roid
   (c) Terman (d) Merrill
9. Raven's progressive matrics (RPM) is an example of
   (a) verbal test (b) nonverbal test
   (c) performance test (d) culture-fair test
10. The concept of mental age was proposed by
   (a) Stern (b) Binet
   (c) Thorndike (d) Guilford
11. Test of creativity measures
   (a) divergent thinking (b) convergent thinking
   (c) both of them (d) none of these
12. Which is not measured by General Aptitude Test Battery (GATB)?
   (a) Numerical aptitude (b) Clerical perception
   (c) Finger dexterity (d) Abstract reasoning

Answers

1. (c) 2. (d) 3. (a) 4. (b) 5. (b) 6. (a) 7. (c)
8. (b) 9. (b) 10. (b) 11. (a) 12. (d)

10. Measurement of Personality

1. The first personality inventory was developed in 1912 by
   (a) Woodworth (b) Guilford
   (c) Galton (d) Kaplan
2. The famous Minnesota Multiphasic Personality Inventory (MMPI) has been developed following
   (a) logical content strategy (b) criterion-group strategy
   (c) factor-analytic strategy (d) mixed strategy
3. MMPI-2 contains
   (a) 550 items (b) 567 items
   (c) 560 items (d) 576 items
4. Which of the following has not been included in MMPI-2?
   (a) Hysteria (b) Paranoia
   (c) Depression (d) Phobia
5. Cattell's 16 PF inventory is based upon
   (a) factor-analytic strategy (b) logical content strategy
   (c) criterion-group strategy (d) mixed strategy
6. Mohan has unconscious resentment towards his father. Which test might best detect this?
   (a) MMPI-2 (b) Rorschach test
   (c) TAT (d) Guilford–Zimmerman Temperment Survey
7. In Catell's 16 PF test, the symbol Q_2 indicates
   (a) imagination (b) impulsivity
   (c) self-sufficiency (d) dominance

Scanned with CamScanner

8. Edwards Personal Preference Schedule measures
   (a) various types of interests
   (b) various types of needs
   (c) various types of traits
   (d) various types of attitudes
9. Which is not assessed by EPQ?
   (a) Psychoticism
   (b) Extraversion
   (c) Neuroticism
   (d) Hypochondriasis
10. Researches have shown that antisocial personality and schizoid personality often score high on ...... of EPQ.
    (a) P scale and E scale
    (b) P scale and N scale
    (c) N scale only
    (d) P scale only
11. Likert scale of attitude measurement is one type of
    (a) nondisguised nonstructured test
    (b) nondisguised structured test
    (c) disguised structured test
    (d) disguised nonstructured test
12. Which is not assessed by the NEO-Personality Inventory (NEO-PI-R)?
    (a) Openness to experience
    (b) Agreeableness
    (c) Neuroticism
    (d) Zestfulness

Answers

1. (a) 2. (b) 3. (b) 4. (d) 5. (a) 6. (c) 7. (c)
8. (b) 9. (d) 10. (d) 11. (b) 12. (d)

11. Projective Techniques

1. In modern time, first projective test was developed by
   (a) Galton
   (b) Frank
   (c) Rorschach
   (d) Morgan
2. Rorschach test measures
   (a) nonintellectual traits
   (b) intellectual traits
   (c) both (a) and (b)
   (d) neither (a) nor (b)
3. Comprehensive Scoring System (CSS) of Rorschach responses was developed by
   (a) Klopfer
   (b) Rapaport
   (c) Exner
   (d) Beck
4. Which of the following can't be included in the categories of classification of responses of Rorschach test?
   (a) Location
   (b) Determinants
   (c) Content
   (d) Context
5. What is the popular response for card No. VI of Rorschach test?
   (a) Bat
   (b) Animal skin
   (c) Human figure
   (d) Fish
6. What does *S* response in Rorschach test indicate?
   (a) Negativism
   (b) Ability to perceive and read clearly
   (c) Concentration
   (d) None of these
7. In Holtzman Inkblot test, how many cards are there?
   (a) 40 cards
   (b) 30 cards
   (c) 45 cards
   (d) 55 cards
8. In TAT, the maximum number of cards to be administered to any one person is
   (a) 20
   (b) 19
   (c) 30
   (d) 25

Scanned with CamScanner

9. The Children Apperception Test (CAT) is used for children
   (a) below 10 years (b) in between 10 to 12 years
   (c) above 12 years (d) in between 13 to 15 years
10. Machover Draw-a-Person (D-A-P) test belongs to the category of
   (a) verbal techniques (b) expressive techniques
   (c) pictorial techniques (d) constructive techniques
11. Which of the following is measured by TAT?
   (a) Attitudes (b) Needs
   (c) Prejudices (d) None of these
12. Which is not a characteristic of projective tests?
   (a) Projective tests are based upon projective hypothesis.
   (b) Projective tests are relatively unstructured.
   (c) Projective tests usually employ a global approach.
   (d) Interpretation of responses of projective tests is influenced by nonpsychoanalytic thinking.

Answers

1. (a) 2. (c) 3. (c) 4. (d) 5. (b) 6. (a) 7. (c)
8. (a) 9. (a) 10. (b) 11. (b) 12. (d)

12. Techniques of Observation and Data Collection

1. Which is not a characteristic of a good questionnaire?
   (a) The questionnaire should be concerned with general topics.
   (b) The questionnaire should be as far as possible short.
   (c) Directions and wordings of the questions should be simple.
   (d) Questions should be presented from general to specific responses.
2. What is the correct major point of distinction between questionnaire and interview schedule?
   (a) Questionnaire should be a short one whereas an interview schedule should be a long one.
   (b) Questionnaires contain questions, which are read and answered by respondents themselves whereas interview schedules contain questions to be read by the researcher before respondents who accordingly respond after listening.
   (c) Questionnaires are prepared by the researcher whereas interview schedules are prepared by respondents.
   (d) Questionnaire takes more time as compared to interview schedule in collecting data
3. The success of interview is dependent upon
   (a) accessibility (b) cognition
   (c) motivation (d) all of these
4. The interview based upon preframed questions to be asked is technically known as
   (a) formal interview (b) informal interview
   (c) semi-formal interview (d) semi-informal interview
5. Which is not a source of error in interview?
   (a) Direction of interview (b) Attitude of interviewer
   (c) Attitude of the interviewee (d) Lack of warmth in the situation of interview
6. Which can't be a purpose of content analysis?
   (a) To analyze the different types of errors in students work
   (b) To find out the relative importance of some topics as problems
   (c) To explain the possible causal factors related to some outcome or event
   (d) To analyze about the needs and conflicts of the writers or the person concerned

Scanned with CamScanner

7. In naturalistic observation, how is the problem of reactivity dealt with?
   (a) The observer becomes a member of the group.
   (b) Observers are kept hidden or the participants are habituated to the presence of observers.
   (c) Participants give informed consent.
   (d) Participants never know they are being observed.
8. Which is not the correct point of difference between participant observation and nonparticipant observation?
   (a) In participant observation the observer actively participates in the activities of the group whereas in the nonparticipant observation, active participation by the observer does not take place.
   (b) Participant observation is usually unstructured whereas nonparticipant observation is usually structured.
   (c) In participant observation the identity of the observer is usually hidden but in nonparticipant observation, the observer is usually known to the persons being observed.
   (d) In participant observation, the number of participants is larger than the number of participants in nonparticipant observation.
9. If you prefer to study the verbalized attitudes of the participants by using a comparative rating method, your ideal choice will be for
   (a) forced-choice scale
   (b) Q-sort technique
   (c) Likert scale
   (d) none of these
10. Semantic differential scale mainly measures
   (a) denotative fact of meaning
   (b) connotative fact of meaning
   (c) association
   (d) none of these
11. Which of the following can't be properly studied by sociometric method?
   (a) Choices of the individuals
   (b) Communication pattern
   (c) Interaction pattern
   (d) Motivational pattern
12. When the rater is predominantly influenced by a single favourable or unfavourable trait of the ratee and accordingly rates him, this constitutes the example of
   (a) error of leninency
   (b) halo error
   (c) logical error
   (d) error of severity

Answers

1. (a) 2. (d) 3. (d) 4. (a) 5. (d) 6. (d) 7. (b)
8. (d) 9. (b) 10. (b) 11. (d) 12. (b)

13. Scaling Techniques

1. In psychophysical scale, the term PSE stands for
   (a) point of subjective equality
   (b) point of subject equality
   (c) point of suspect equality
   (d) point of stupid equality
2. The error of habituation and the error of anticipation are associated with
   (a) method of average error
   (b) method of limits
   (c) method of constant stimuli
   (d) none of these
3. The relationship between the size of the standard stimulus and the size of just noticeable difference (JND) is technically known as
   (a) Fechner's law
   (b) Weber' law
   (c) Steven's power law
   (d) law of JND

Scanned with CamScanner

4. Which of the following can be used both as a method of psychophysical scaling as well as a method of psychological scaling?
   (a) Method of category scaling
   (b) Method of magnitude estimation
   (c) Method of paired comparison
   (d) None of these
5. In the method of equal-appearing interval, the scale value of the item is indicated by
   (a) median
   (b) standard deviation
   (c) quartile deviation
   (d) none of these
6. Which of the following can't be considered as a disadvantage of method of equal appearing interval?
   (a) The judges don't keep the intervals equal.
   (b) The attitude of the judges tends to influence the sorting of statements.
   (c) In this method there is no objective basis for selecting the most discriminating items.
   (d) The judges remain dissatisfied with sorting of all statements in only 11-point.
7. In which scaling technique are items selected on the basis of rigorous procedure of item analysis?
   (a) Method of summated ratings
   (b) Method of equal appearing interval
   (c) Cumulative scale
   (d) None of these
8. According to Guttman if the coefficient of reproducibility is between 0.85 to 0.90, it is said that
   (a) quasi-cumulative scale is existing
   (b) a perfect cumulative scale is existing
   (c) a perfect social distance scale is existing
   (d) none of these
9. Under which scaling method is there provision for 5-point response options?
   (a) Method of summated rating
   (b) Method of cumulative scaling
   (c) Method of successive categoring
   (d) Method of equal appearing interval
10. Law of comparative judgement was formulated by
   (a) Likert
   (b) Thurstone
   (c) Stevens
   (d) Fechner

Answers

1. (a) 2. (b) 3. (b) 4. (b) 5. (a) 6. (d) 7. (a)
8. (a) 9. (a) 10. (b)

14. Sampling

1. Which is a correct statement?
   (a) Sample is a smaller representation of population.
   (b) Sample is a bigger representation of population.
   (c) Every sample contains many types of error.
   (d) Decision to sample is not influenced by the size of the population.
2. When all elements in a given population have an equal and independent chance of being included in the sample, the process is referred to as
   (a) good sampling
   (b) random sampling
   (c) nonrandom sampling
   (d) none of these
3. Which of the following is a correct statement?
   (a) Systematic sampling is one type of nonprobability sampling.
   (b) Systematic sampling does not help in making correct and scientific judgement.

Scanned with CamScanner

(c) Systematic sampling is partially probability sampling and partially nonprobability sampling.
(d) Systematic sampling takes into consideration all persons.

4. Quota sampling greatly resembles
   (a) area sampling (b) stratified random sampling
   (c) purposive sampling (d) snowball sampling

5. Difference between parameter and statistic is called as
   (a) sampling error (b) error of sampling distribution
   (c) error score (d) standard error of difference

6. List of items from which sampling is to be finally drawn is called as
   (a) sampling units (b) sampling list
   (c) sampling frame (d) none of these

7. Which is not a characteristic of a good sampling design?
   (a) The sample design must provide a sample, which would be a true representative of the population.
   (b) Sampling design must control all errors except sample error.
   (c) Sampling design must be usable in light of the funds available for the research study.
   (d) The sample design must be such that the results of the sample study can be applied in general.

8. Another name of accidental or incidental sampling is
   (a) convenience sampling (b) purposive sampling
   (c) judgemental sampling (d) inferential sampling

9. When sampling is drawn by selecting every *n*th person from a predetermined list of individuals, the technique is called as
   (a) dense sampling (b) saturation sampling
   (c) snowball sampling (d) systematic sampling

10. The numerical value based upon sample is called ...... whereas the numerical value based upon population is known as ....
    (a) statistic, parameter (b) parameter, statistic
    (c) statistical error, parametric error (d) parametric error, statistical error

11. What is the advantage of cluster sampling over simple random sampling and stratified sampling?
    (a) Cluster sampling allows the researcher to represent subgroups of the population accurately.
    (b) In cluster sampling, it is not essential to begin with a complete list of the population.
    (c) Cluster sampling is the only sampling that could be called as convenience sampling.
    (d) Cluster sampling improves probability sampling.

12. If the population is extremely large, which of the following sampling becomes impractical?
    (a) Stratified random sampling (b) Random sampling
    (c) Cluster sampling (d) Snowball sampling

Answers

1. (a) 2. (b) 3. (c) 4. (b) 5. (a) 6. (c) 7. (b)
8. (a) 9. (d) 10. (a) 11. (b) 12. (b)

Scanned with CamScanner

15. Social Scientific Research

1. Which is not a relevant statement?
   (a) Research is always directed towards the solution of problem.
   (b) Research is characterized by systematic, objective and logical procedures.
   (c) Research is replicable.
   (d) Research needs good communication skills on the part of the researchers.

2. Compared with field research, which of the following is true about laboratory research?
   (a) Informed consent is difficult to be gained in laboratory research than in field research.
   (b) Laboratory research achieves greater mundane realism.
   (c) For achieving experimental realism, the study needs to be conducted in the laboratory.
   (d) Laboratory research achieves a greater degree of control over the different conditions of the experiment.

3. One study compares three-month and five-month-old children on a perceptual task. Another study compares four groups of children of age 3, 5, 7 and 9 months using the same task. The second study
   (a) is a good example of replication and extension
   (b) is redundant because there is no need to include the first two groups
   (c) is an example of diverging operation
   (d) has higher degree of experimental reality than the first one

4. Which research has greater degree of internal validity?
   (a) Laboratory experiment research (b) Field experiment research
   (c) Field study research (d) Ex post facto research

5. About a year prior to retirement, 150 people volunteer for a programme to reduce high blood pressure. Two years after retirement, all 150 people remain in the programme and their blood pressure was found to be down. It is probably difficult to interpret these data because of what type of threat to internal validity?
   (a) History (b) Instrumentation
   (c) Testing (d) Maturation

6. In an experimental study of memory, the researcher hypothesizes that using visual imagery enhances recall of words. Some participants were told to create images and some were told to repeat each word two times. Since forming images takes longer than repeating, those in the imagery group are shown the words at a slower presentation rate (5 seconds per word) than those in the repetition group (3 seconds per word). Which of the following is correct about this study?
   (a) It has internal validity but lacks external validity.
   (b) The rate of the presentation is the independent variable.
   (c) Presentation rate is confounded with instructional independent variable.
   (d) It is a good study with an instructional variable as the manipulated independent variable.

7. For an experimental group in a programme to reduce anxiety, the mean pretest score is 89 (maximum = 100) and the mean post-test score is significantly lowered, that is, 80. This change is probably due to
   (a) maturation (b) regression
   (c) programme effectiveness (d) any of the above (or some combination)

8. Which is not a proper method of controlling threats to reliability and validity in research?
   (a) Reliable manipulation (b) Pilot studies
   (c) Defence against diffusion of treatment
   (d) Statistical analysis

Scanned with CamScanner

9. The major advantage of interview over written and phone surveys is that interview
   (a) is less expensive (b) avoids problem of prejudice and bias
   (c) yields information in detail (d) never has sampling problem
10. Which is true regarding case study as a method?
    (a) It is a correct and reasonable choice for dealing with rare individuals.
    (b) Experimenter bias is not likely to be a problem.
    (c) Case study allows to draw conclusion based on cause and effect.
    (d) Case study allows in-depth studies of individual persons.
11. After the general election is over in the country, a social psychologist interviewed 400 voters to study their attitudes towards the winning party. This becomes an example of
    (a) correlational study (b) field experiment
    (c) ex post facto research (d) voting behaviour experiment
12. The common description, "I-wonder-what-would-happen-if-I-did-this" relates to
    (a) crucial experiment (b) confirmatory experiment
    (c) exploratory experiment (d) pilot experiment

Answers

1. (d) 2. (d) 3. (a) 4. (a) 5. (a) 6. (c) 7. (d)
8. (d) 9. (c) 10. (a) 11. (c) 12. (c)

16. Single-subject Experimental Research and Small *N* Research

1. Why do we prefer an A-B-A-B design to an A-B-A design?
   (a) Because it evaluates the treatment effect twice rather than once
   (b) Because it includes a withdrawal of treatment
   (c) Because it compares contingent with noncontingent reinforcement
   (d) Because it is more parsimonious
2. When is a multiple baseline design preferred over an A-B-A-B design?
   (a) When the purpose is to study more than a single individual
   (b) When withdrawing treatment is not feasible for some reason
   (c) When the purpose is to compare two different treatment strategies in the same participant
   (d) When the researcher feels that the target behaviour can't be researched all at once
3. Which one of the following arguments is not made in favour of using small N design?
   (a) When data for many subjects are averaged, the emerging picture does not reflect the behaviour of any single individual.
   (b) If the precise control of conditions yields orderly behaviour then emphasis should be given on controlling environmental conditions for single subject rather than on achieving statistical controls.
   (c) It is not possible to research on relatively rare psychological phenomenon except by using small N design.
   (d) Because of individual differences, inferential statistics are needed to determine if the independent variable is responsible for some changes in the dependent variable.
4. When the researcher wants to apply treatment in sequence to the same class of behaviour in different participants who are in the same environment, his most preferred design would be
   (a) reversal design (b) multiple-baseline design across subjects
   (c) multiple-baseline design across behaviours
   (d) withdrawal design

Scanned with CamScanner

5. When a researcher wants to determine the status of the subject's behaviour prior to the intervention, he will use
   (a) basephase (b) baseline
   (c) base design (d) multiple-baseline designs

6. Which is not a limitation of single *N* research?
   (a) Lack of external validity (b) Order effects
   (c) Baseline problems (d) Control of experimental situation

7. Which is not a type of withdrawal design?
   (a) A-B-A design (b) A-B-A-B design
   (c) A-B design (d) A_1-B_1, B_2-B_2 design

8. Which is not a good data collection strategy in single-subject research?
   (a) Method of interval recording (b) Frequency measure
   (c) Method of common error (d) Response-specific measure

9. For mentally retarded children, two treatment programmes were started and the researcher wishes to evaluate these programmes. The task was to get the children make their beds properly each morning. One treatment programme used a token economy system in which points can be earned and they can be traded for their reinforcers. The second treatment programme relied on the use of staff attention of a reinforcer, in which the bed making is reinforced by giving the child extra attention. For evaluating the effectiveness of the strategies, which would be the best design?
   (a) Withdrawal design (b) A-A_1-B-A_1-B design
   (c) A-B-C design (d) Alternating treatment design

10. For changing the self-stimulation behaviour of an autistic child, a reinforcement plan is put into effect in three different settings but is staggered so that contingencies are introduced in one setting at a time. Here the researcher is using
    (a) multiple-baseline design (b) reversal design
    (c) A-B-A design (d) A-B-A-B design

Answers

1. (a) 2. (b) 3. (d) 4. (b) 5. (b) 6. (d) 7. (d)
8. (c) 9. (d) 10. (a)

17. Historical Research

1. Which is a correct statement?
   (a) Historical research is a philosophical investigation of the past events.
   (b) Historical research is the critical investigation of events and experiences of the past.
   (c) In historiography, the researcher pays attention to only internal criticism.
   (d) In historiography, the investigator pays attention to only external criticism.

2. Which is not a correct statement?
   (a) Historiography is relevant because all the data used in social sciences are drawn from the recorded experiences of the past.
   (b) The thinking of all workers in the social sciences and education is determined by the historical circumstances of their lives.
   (c) All policies respecting human affairs in social sciences and humanities involve interpretation of the past.
   (d) Social scientists can't interpret any of the behaviour without direct reference to the past.

Scanned with CamScanner

3. Oral testimony is one of the
   (a) secondary source of data
   (b) primary source of data
   (c) convincing source of data
   (d) none of these
4. Which is a correct statement?
   (a) Historical research is quantitative in nature.
   (b) Historical research is qualitative in nature.
   (c) Historical research is both quantitative and qualitative in nature.
   (d) Historical research is neither quantitative nor qualitative in nature.
5. Which is an eyewitness account in historical research?
   (a) Relics or remain
   (b) Oral testimony
   (c) Documents
   (d) All of these

Answers

1. (b) 2. (d) 3. (b) 4. (c) 5. (d)

18. The Problem and the Hypothesis

1. Which is not a characteristic of research problem?
   (a) A problem statement is usually written in an affirmative form.
   (b) A problem statement expresses the relationship between two or more than two variables.
   (c) A problem statement should be testable by empirical method.
   (d) A problem statement should avoid moral or ethical judgement.
2. According to P V Young, which is not a source of locating a research problem?
   (a) Documentary source
   (b) Personal source
   (c) Library source
   (d) Mass-media source
3. A problem is manifested in the following ways except
   (a) a noticeable gap in the results of investigations
   (b) contradictory results of investigations
   (c) isolated fact in the form of unexplained information
   (d) well-settled fact in the area
4. 'Is there a divinity that shapes our ends?' is an example of
   (a) solvable problem
   (b) partly solvable problem
   (c) unsolvable problem
   (d) partly unsolvable problem
5. Which is not a characteristic of good hypothesis?
   (a) Hypothesis should be tenable.
   (b) Hypothesis should be economical and parsimonious.
   (c) Hypothesis should be specific in scope.
   (d) Hypothesis should have logical unity.
6. A researcher often experiences difficulties in formulating hypothesis. Which of the following can't be included in the list of difficulties?
   (a) Absence of knowledge
   (b) Lack of ability to utilize the knowledge of theoretical framework
   (c) Lack of knowledge about scientific research technique
   (d) Lack of proper training
7. Which is not a correct statement?
   (a) A research hypothesis is derived from the researcher's theory.
   (b) Null hypothesis is no-effect or negative hypothesis.
   (c) Research hypothesis is the reverse of alternative hypothesis.
   (d) A universal hypothesis is one in which the stated relationship holds good for all the levels or values of variable.

Scanned with CamScanner

8. Which is not a correct function of hypothesis?
   (a) Hypothesis suggests various theories.
   (b) Hypothesis tests various theories.
   (c) Hypothesis tends to describe social phenomena.
   (d) Hypothesis provides basis for evaluating supernatural phenomena.

9. A hypothesis is generated from various sources except
   (a) from various opinions, experiences and observations
   (b) by retesting of hypothesis previously tested
   (c) from existing research itself
   (d) from observing the behaviours of others.

10. A laboratory experimenter, in general, prefers to formulate a hypothesis called as
    (a) existential hypothesis (b) universal hypothesis
    (c) causal hypothesis (d) descriptive hypothesis

Answers

1. (a) 2. (d) 3. (d) 4. (c) 5. (c) 6. (d) 7. (c)
8. (d) 9. (d) 10. (c)

19. Reviewing the Literature

1. Which can't be a purpose of review of literature?
   (a) Avoidance of repetition
   (b) Determining relevance of hypothesis
   (c) Identifying variables relevant for research
   (d) Synthesis of prior works

2. All of the following are the sources of the review except
   (a) books (b) journals
   (c) internet (d) television

3. A literature, either published or unpublished, bearing no ISBN or ISSN is technically called as
   (a) white literature (b) grey literature
   (c) black literature (d) none of these

4. An ideal index card for reviewing and abstracting generally contains
   (a) 150 words approximately (b) 160 words approximately
   (c) 200 words approximately (d) 250 words approximately

5. In psychological research, a very popular psychological abstract was first published in
   (a) 1927 (b) 1937
   (c) 1910 (d) 1920

Answers

1. (b) 2. (d) 3. (b) 4. (a) 5. (a)

20. Variables

1. If the researcher wants to study the impact of the level of illumination upon reading ability of a person, the dependent variable here will be
   (a) reading ability (b) level of illumination
   (c) researcher himself (d) none of these

Scanned with CamScanner

2. The variable which is directly manipulated by the experimenter is known as
   (a) type-*S* independent variable (b) type-*E* independent variable
   (c) type-*S* dependent variable (d) type-*E* dependent variable
3. Sometimes the researcher manipulates a variable only for the sake of controlling its unwanted effect upon the experimental results. Such variable is known as
   (a) extraneous variable (b) irrelevant variable
   (c) controlled independent variable (d) none of these
4. Members of the family constitute the example of
   (a) quantitative variable (b) categorical variable
   (c) continuous variable (d) none of these
5. Which is not a correct method of controlling extraneous variable?
   (a) Randomization (b) Balancing
   (c) Constancy of conditions (d) Constancy of variables
6. Which does not produce demand characteristics?
   (a) Experimenter (b) Subject
   (c) Measurement procedures (d) Technique of elimination
7. A variable which is not manipulated but measured by the experimenter is known as
   (a) intervening variable (b) attribute variable
   (c) active variable (d) none of these
8. When a researcher uses statistics properly and draws the appropriate conclusions from the statistical analysis, the process is named as
   (a) statistical validity (b) statistical conclusion validity
   (c) statistical answer validity (d) validity of inferential statistics
9. Which is not the appropriate method of manipulating independent variables?
   (a) By manipulating social setting in which confederates are used
   (b) By manipulating likely information to be given to the subjects
   (c) By manipulating intervening variables
   (d) By manipulating experimenter's role
10. Which of the following is a continuous variable?
    (a) Reaction time (b) Religion
    (c) Race (d) Caste

Answers

1. (a) 2. (b) 3. (a) 4. (b) 5. (d) 6. (d) 7. (b)
8. (b) 9. (d) 10. (a)

21. Research Design

1. The major advantage of between-subjects (or groups) design over within-subjects (or group) design is that the former design
   (a) avoids the problem of equivalent groups
   (b) reduces the extent of error variance between conditions
   (c) requires fewer subjects
   (d) avoids the problem of sequence effects
2. When asymmetric transfer effect occurs
   (a) the matching must be used as true method for creating equivalent groups
   (b) complete counterbalancing must be used
   (c) partial rather than complete counterbalancing must be used
   (d) counterbalancing may not prove fruitful in eliminating the sequence effects

Scanned with CamScanner

3. As a technique of creating equivalent groups, the researcher prefers matching over random assignment
   (a) wherever a measurable extraneous variable is known to correlate with the dependent variable
   (b) every time when a larger number of participants is available
   (c) whenever potential confounding effect exists
   (d) in all the above conditions
4. The experimenter expectancy effects
   (a) have never been replicated after Rosenthal's research, so they are not real problems
   (b) can be easily reduced by automating the procedure as much as possible
   (c) are often found in research utilizing human subjects but not in researches utilizing animal subjects
   (d) does not occur if the participants are unaware of the hypothesis being tested
5. Block randomization is often used
   (a) for counterbalancing when the participants are tested more than once per condition in a within-subjects design
   (b) for achieving random assignment while ensuring that an equal number of participants are assigned to each group
   (c) none of the above
   (d) for both a and b
6. Which design is illustrated by the case study comparing average children and gifted children on social and emotional problem-solving task?
   (a) Independent groups
   (b) Matched group
   (c) Within-subjects, multilevel
   (d) Nonequivalent groups
7. What do all repeated-measures designs (single factor) have in common?
   (a) Every participant will be tested in each of the conditions of the study.
   (b) For creating equivalent groups, matching will be preferred.
   (c) They will always have a control group.
   (d) The participants will be tested in each condition of the study and will always be tested more than once per condition.
8. A researcher predicts that an extrovert will do better on a problem-solving task with an audience rather than when alone. Introverts, however, are expected to perform better in alone condition than with audience. Here the researcher is
   (a) using a mixed factorial design
   (b) expecting an interaction effect to occur
   (c) using two manipulated variables
   (d) trying to predict a main effect for the audience variable
9. A 2 × 2 mixed factorial design includes
   (a) one manipulated factor and one nonmanipulated factor
   (b) a within-subjects variable and a repeated-measures variable
   (c) a between-subjects factor and a repeated-measures variable
   (d) a two-subject variable and a three-subject variable
10. In a maze-learning experiment the experimenter took 60 rats, of which 30 rats were tested with the light-on condition and 30 rats were tested with light-off condition. Also within each of the two groups, 10 rats were given food immediately after they reached the goal, 10 others were reinforced 8 seconds after they reached the goal and the remaining 10 were fed 12 seconds after reaching the goal. What is your idea about this design?
    (a) It is 2 × 2 independent groups design.

Scanned with CamScanner

(b) It is 2 × 3 mixed design.
(c) Six different conditions are being tested.
(d) Both independent variables are subject factors.

11. A $2 \times 3 \times 5$ design has
    (a) dependent variables with two, three and five levels respectively
    (b) 30 different independent variables
    (c) 30 different conditions
    (d) 10 different independent variables

12. A 3×3 mixed factorial design uses five participants in each cell of the design. How many participants are needed for this purpose?
    (a) 9
    (b) 27
    (c) 45
    (d) 15

13. In correlational studies, the extraneous variables are not controlled. Consequently, a problem is created. Which of the following is that problem?
    (a) Range restriction problem
    (b) Third variable problem
    (c) Directionality problem
    (d) Problems relating to measurement of dependent value

14. A cross-lagged panel technique is sometimes used to solve problems relating to
    (a) directionality
    (b) third variable
    (c) negative correlation
    (d) positive correlation

15. In an unequivalent control group design
    (a) on the basis of results the researcher is able to evaluate the linear trends
    (b) it is highly essential that the scores of the both groups must be same on the pretest
    (c) the most important measure is usually the amount of change between pretest and posttest
    (d) threats to internal validity by selection and history can be easily ruled out because of the control group

16. Suppose the Bihar state wants to implement strictly the law relating to wearing helmets by riders on motorcycles because of a recorded number of head injuries. Then which of the following will be true?
    (a) Comparison with a similar state that does not have the law would help assess potential regression effects.
    (b) The number of head injuries is likely to decline the following year.
    (c) Evaluation calls for an interrupted time series design.
    (d) All of the above

Answers

1. (d) 2. (b) 3. (a) 4. (b) 5. (d) 6. (d) 7. (a)
8. (b) 9. (c) 10. (c) 11. (c) 12. (d) 13. (b) 14. (a)
15. (c) 16. (d)

22. Qualitative Research

1. Which is not true about qualitative research?
    (a) Qualitative research is based upon phenomenological paradigm.
    (b) Qualitative researchers use highly standardized instruments at the outset.
    (c) In qualitative research, most of the analysis is done with words.
    (d) In qualitative research, the researcher aims to gain a holistic view of the context under study.

Scanned with CamScanner

2. Which of the following questions is not answered by qualitative research?
   - (a) What does the qualitative researcher actually do?
   - (b) What does the researcher want to understand at the end of the study?
   - (c) What are the contexts in which the study is done?
   - (d) What rigorous research design should be adopted in the context?
3. Which method of sampling is generally preferred by a qualitative researcher?
   - (a) Quota sampling
   - (b) Purposive sampling
   - (c) Snowball sampling
   - (d) Cluster sampling
4. When a researcher explores and analyzes the data without prior hypothesis, it is called as
   - (a) context sensitivity
   - (b) inductive analysis
   - (b) deductive analysis
   - (d) hypothetical analysis
5. In which of the following the researcher tries to answer the question: How do persons try to accomplish their goals through specific behaviours in specific environment?
   - (a) Ethnomethodology
   - (b) Heuristic enquiry
   - (c) System theory
   - (d) Ecological psychology
6. When the researcher tries to generate theory from data, the process is referred to as
   - (a) ethnography
   - (b) grounded theory
   - (c) secondary theory
   - (d) primary theory
7. Which is not used by the qualitative researcher?
   - (a) Grounded theory
   - (b) Restrospective studies
   - (c) Content analysis
   - (d) Factor analysis
8. Which is not a characteristic of case study?
   - (a) A case is a bounded system.
   - (b) In case study, no attempt is made to preserve the wholeness and integrity of the case.
   - (c) In case study, multiple data collection methods are used in naturalistic setting.
   - (d) The case is case of something.
9. When the researcher takes cases, groups or materials according to their relevance for the theory that is developed and on the background of what is already the state of knowledge after collecting and analyzing a certain number of cases, the procedure is known as
   - (a) grounded theory
   - (b) theoretical sampling
   - (c) typical case sampling
   - (d) sensitive case sampling
10. Which is not a component of qualitative data analysis according to Miles and Huberman?
    - (a) Data display
    - (b) Data reduction
    - (c) Drawing conclusion
    - (d) Interpreting the results

Answers

1. (b) 2. (d) 3. (b) 4. (b) 5. (d) 6. (b) 7. (d)
8. (b) 9. (b) 10. (d)

23. Carrying out Statistical Analysis

1. Which is not a characteristic of normal curve?
   - (a) A normal curve is continuous.
   - (b) A normal curve is asymptotic to the x-axis.
   - (c) A normal curve is usually bimodal.
   - (d) A normal curve is symmetrical.

Scanned with CamScanner

2. In normal curve at $\pm 2\sigma$, how many percentage of cases lie?
   (a) 95.44%
   (b) 95.00%
   (c) 84.44%
   (d) 68.26%

3. When the researcher estimates parameters using single value, these estimates are known as
   (a) interval estimation
   (b) point estimation
   (c) confidence estimation
   (d) none of these

4. When null hypothesis is false and the researcher fails to reject this hypothesis, this constitutes the example of
   (a) Type II error
   (b) Type I error
   (c) Type III error
   (d) none of these

5. Which one measures the degree to which the relation between two variables X and Y is one-directional or monotonic?
   (a) Pearson correlation
   (b) Kendall's partial correlation
   (c) Coefficient of contingency
   (d) Spearman correlation

6. What will happen when the researcher reduces alpha level?
   (a) The risk of Type I error is reduced.
   (b) The risk of Type II error is reduced.
   (c) The risk of Type I error is enhanced.
   (d) The risk of Type II error is enhanced.

7. Which is not used as post hoc test?
   (a) Tukey's HSD
   (b) Duncan multiple range test
   (c) Neuman–Keuls test
   (d) Chi-square test

8. Which is not the correct assumptions of ANCOVA?
   (a) Treatment groups are selected at random from the same population.
   (b) X-scores (covariate) are affected by the treatment.
   (c) Within-group variances should be approximately equal.
   (d) The dependent variable that is under measurement should be normally distributed in the population.

9. Which is the measure of absence of relationship between X and Y?
   (a) Coefficient of alienation
   (b) Coefficient of determination
   (c) Coefficient of distribution
   (d) Coefficient of linear regression

10. In regression equation, the differences between a score and its expected value according to theoretical model is known as
    (a) third variable explanation
    (b) residual
    (c) standard error of estimate
    (d) none of these

11. In regression analysis, the sum of residuals is always equal to
    (a) +2
    (b) −2
    (c) +1
    (d) 0

12. In general regression equation, the slope of the regression line is indicated by
    (a) $y - bx$
    (b) $y = a + bx$
    (c) only b
    (d) only a

Answers

1. (c) 2. (a) 3. (b) 4. (a) 5. (d) 6. (a) 7. (d)
8. (b) 9. (a) 10. (b) 11. (d) 12. (c)

Scanned with CamScanner

24. Writing a Research Report and a Research Proposal

1. In writing a research report, what is placed immediately before introduction?
   (a) Methods of the study (b) Hypothesis
   (c) Abstract (d) None of these
2. Which is a correct statement?
   (a) Bibliography and references are the same thing.
   (b) Bibliography contains a detailed description of the history of the researcher.
   (c) References contain the details of the research reports of only those authors whose work has been included.
   (d) According to APA, references are usually placed at the end of each chapter of the research report.
3. In writing research report in the field of psychology and education, we usually follow the guidelines of
   (a) *Publication Manual* of APA (b) Chicago *Manual of Style*
   (c) *A manual for writing of Term papers, theses and dissertations*
   (d) *The MLA Handbook for writers of research papers.*
4. Which portion of the research report ensures that the researcher is familiar with what is already known and what is still unknown and to be verified and tested?
   (a) Introduction (b) Discussion
   (c) Review of literature (d) Conclusion
5. Which section of the research report or research proposal is known as the 'heart' of the research work?
   (a) Introduction (b) Methods
   (c) Results (d) Discussion

Answers

1. (c) 2. (c) 3. (a) 4. (c) 5. (b)

❖

Scanned with CamScanner

APPENDICES

APPENDIX A: MAJOR EVENTS IN THE HISTORY OF MEASUREMENT AND PSYCHOLOGICAL TESTING

206 BC	Chinese start Civil Service Examinations.
1838	Jean Esquirol distinguishes between mental illness and mental retardation.
1862	Wilhelm Wundt uses a calibrated pendulum to measure the "speed of thought."
1866	O Edouard Seguin writes the first major textbook on the assessment and treatment of mental retardation.
1879	Wundt founds the first experimental laboratory in psychology in Leipzig University, Germany, now known as Karl Marx University.
1879	Galton develops the first projective technique—a Word Association Test.
1885	Sir Francis Galton establishes first Mental Testing Centre in London.
1890	James McKeen Cattell uses the term *mental test* in announcing the agenda for his Galtonian test battery.
1896	Emil Kraepelin provides the first comprehensive classification of mental disorders.
1901	Clark Wissler discovers that Cattellian "brass instruments" tests have no correlation with college grades.
1904	Charles Spearman proposes that intelligence consists of a single general factor g and numerous specific factors s_1, s_2, s_3, and so forth.
1904	Karl Pearson formulates the theory of correlation.
1905	Alfred Binet and Theodore Simon invent the first modern intelligence test.
1908	Henry H Goddard translates the Binet–Simon scales from French into English.
1912	Stern introduces the IQ, or intelligence quotient: the mental age divided by chronological age.
1916	Lewis Terman revises the Binet–Simon scales, publishes the Stanford–Binet; revisions appear in 1937, 1960 and 1986.
1917	Robert Yerkes spearheads the development of the Army Alpha and Beta examinations used for testing WWI recruits.
1917	Robert Woodworth develops the Personal Data Sheet, the first personality test.
1920	The Rorschach Inkblot Test is published.
1920	Culture Fair Intelligence Test, a nonverbal measure of fluid intelligence, is conceived by Cattell.
1921	Psychological Corporation—the first major test publisher—is founded by Cattell, Thorndike and Woodworth.
1924	Indian Psychological Association is founded
1926	Florence Goodenough publishes the Draw-A-Man Test.
1926	The first Scholastic Aptitude Test is published by the College Entrance Examination Board.
1927	The first edition of the Strong Vocational Interest Blank is published.
1931	Allport Vernon study of value is released.
1935	The Thematic Apperception Test is published by Morgan and Murray at Harvard University.

Scanned with CamScanner

1936	Edgar Doll publishes the Vinel and Social Maturity Scale for assessment of adaptive behaviour in those with mental retardation.
1938	L L Thurstone proposes that intelligence consists of about seven group factors known as primary mental abilities.
1938	Raven publishes the Raven's Progressive Matrices, a nonverbal test of inductive reasoning intended to measure Spearman's *g* factor.
1938	Lauretta Bender publishes the Bender Visual Motor Gestalt Test, a design-copying test of visual-motor integration.
1938	Oscar Buros publishes the first *Mental Measurements Yearbook.*
1938	Arnold Gesell releases his scale of infant development.
1939	The Wechsler–Bellevue Intelligence Scale is published; revisions are published in 1955 (WAIS), 1981 (WAIS-R), and 1997 (WAIS-III).
1939	The Kuder Preference Record, a forced-choice interest inventory, is published.
1943	The Minnesota Multiphasic Personality Inventory (MMPI) is published.
1947	The famous Differential Aptitude Test is released. Currently its fifth edition has been released in 1992.
1948	Office of Strategic Services (OSS) uses situational techniques for selection of officers.
1949	The Wechsler Intelligence Scale for Children is published; revisions are published in 1974 (WISC-R) and 1991 (WISC-III).
1949	Cattell's 16 PF test is released.
1949	Psychological Research Wing of the Defence Science Organization of India comes into being.
1950	The Rotter Incomplete Sentences Blank is published.
1951	Lee Cronbach introduces coefficient alpha as an index of reliability (internal consistency) for tests and scales.
1952	American Psychiatric Association publishes the *Diagnostic and Statistical Manual (DSM-I).*
1953	Stephenson develops the Q-technique for studying the self-concept and other variables.
1955	National Institute of Mental Health and Neurosciences (NIMHANS) is established in Bangalore, India.
1956	The Halstead–Reitan Test Battery begins to emerge as the premiere test battery in neuropsychology.
1957	C E Osgood describes the semantic differential.
1957	California Psychological Inventory (CPI) is developed to measure dimensions of Normal Personality.
1958	Lawrence Kohlberg publishes the first version of his Moral Judgement Scale; research with it expands until the mid–1980s.
1959	Campbell and Fiske publish a test validation approach known as the multitrait-multimethod matrix.
1963	Raymond Cattell proposes the theory of fluid and crystallized intelligences.
1968	American Psychiatric Association publishes DSM-II.
1969	Nancy Bayley publishes the Bayley Scales of Infant Development (BSID). The revised version (BSID-2) is published in 1993.
1969	Arthur Jensen proposes the genetic hypothesis of African American versus white IQ differences in the *Harvard Educational Review.*

Scanned with CamScanner

1971	George Vaillant popularizes a hierarchy of 18 ego adaptive mechanisms and describes a methodology for their assessment.
1972	First Survey of Research in Psychology in India is published.
1974	Rudolf Moos begins publication of the Social Climate Scales to assess different environments.
1974	Friedman and Rosenman popularize the Type A coronary-prone behaviour pattern; their assessment is interview-based.
1978	Jane Mercer publishes SOMPA (System of Multicultural Pluralistic Assessment), a test battery designed to reduce cultural discrimination.
1985	NEO-PI was published by Costa and McCrae. Its revised edition is released in 1992.
1985	The American Psychological Association (APA), the American Educational Research Association (AERA) and the National Council on Measurement Education (NCME) jointly publish the influential *Standards for Educational and Psychological Testing.* Its version appears in 1999.
1985	Sparrow and others publish the Vineland Adaptive Behavior Scales, a revision of the pathbreaking 1936 Vineland Social Maturity Scale.
1987	American Psychiatric Association publishes DSM-III-R.
1988	*Psychology in India* (Three volumes) is released.
1989	The Lake Wobegon Effect is noted: Virtually all states of the union claim that their achievement levels are above average.
1989	National Academy of Psychology (NAOP), India, is founded.
1989	Luria–Nebraska Neuro Psychological Battery is released. Its third edition is released in 1989.
1989	The Minnesota Multiphasic Personality Inventory-2 is published.
1992	American Psychological Association publishes a revised *Ethical Principles of Psychologists and Code of Conduct* (*American Psychologist*, December 1992).
1994	American Psychiatric Association publishes DSM-IV.
1994	Herrnstein and Murray revive the race and IQ heritability debate in *The Bell Curve.*
1999	APA and other groups publish revised *Standards for Educational and Psychological Testing.*
2000	DSM IV (TR) is published.
2000	*Psychology in India Revisited* (Three volumes) is released.
2003	Fifth Edition of Stanford–Binet Intelligence test (SB5) is released.
2003	New revision of APA Ethical Principles of Psychologists and code of conduct goes into effect.
2009–11	*Psychology in India* (Four volumes) is released.
2013	American Psychiatric Association Publishes DSM-V.

Scanned with CamScanner

GLOSSARY

A-B design: A type of small N design in which a baseline period (that is, A) is followed by a treatment period (that is, B).

A-B-A design: A type of small N design in which baseline period (A) is followed by a treatment period (B), followed by a period in which the treatment is either reversed or withdrawn (second A).

ABBA counterbalancing: A technique for controlling practice effects in the complete within-subjects design, in which conditions in one sequence are presented and then followed by the opposite of the same sequence.

A-B-C-B design: A small N design that makes comparison of contingent reinforcement (B) with noncontingent reinforcement (C) and subsequently, the researcher is allowed to separate the effects of reinforcers and contingency.

Abscissa: The horizontal axis (or x-axis) in a graph.

Absolute threshold: The point at which the subject or observer is able to detect a physical stimulus about 50 per cent of the time.

Accidental samples: A type of nonprobability sample which is generally low in representativeness and is obtained when easy availability and willingness to respond are the major considerations in selecting the respondents.

Acquiescence: Tendency to endorse or agree with test items.

Adjusting for: Same as partialling out.

Adjusted mean: A mean of a group after adjusting for or a partialling out the effect of a covariate in ANCOVA.

Age differentiation: A discrimination based on the fact that older children possess greater capabilities than younger children.

Age norms: A type of norms that displays the level of test performance for each separate age group in the normative sample.

Age scale: A scale or test in which items are grouped according to age level (for example, Binet test).

Alpha level: The significance level; probability of committing type I error.

Alternative hypothesis (H_1): The researcher's hypothesis about the outcome of a study.

Alternate-forms reliability: A type of reliability in which alternate forms of the same test are developed, given to a group of subjects and then they are correlated.

Analysis of covariance (ANCOVA): A method of statistical analysis based on ANOVA that uses statistical control for removing the effect of an extraneous variable, assumed or known to be correlated with dependent variable and tends to analyze the portion of the variance in the dependent variable explained by one or more independent variables excluding the extraneous variables.

Analysis of variance (ANOVA): A statistical test appropriate for analyzing interval data obtained from between-subject and within-subject experiment designs. The F test is computed for analyzing total variance into different components to answer empirical questions like whether two or more groups in an experiment significantly differ or not.

A posterior comparison: Comparison not planned in advance for investigating specific hypothesis concerning parameters of the population.

Applied research: Such a research seeks knowledge that will improve the present situation because here, abstract principles are applied in real-world settings.

A priori or planned comparison: Comparison planned in advance to investigate specific hypothesis relating to parameters of the population.

Archival data: A type of data that is based upon records or documents concerning individuals, institutions, governments and other groups; commonly used as an alternative or in conjunction with other research techniques.

Archival research: A descriptive method of research in which already existing records are examined to test some research hypothesis.

Assumption: Conditions that must be met before certain conclusions are drawn.

Asymmetric measure of association: A measure of the one-way effect of one variable upon another.

Asymmetrical transfer: Such a differential transfer occurs in the repeated-treatments design where transfer from one condition to a second is different from the second to the first condition.

Attrition: A threat to internal validity of test that occurs when participants fail to complete a study usually found in longitudinal studies. Those finishing the study may not be equivalent to those who started it.

Balanced Latin square: A counterbalancing strategy in which each condition is preceded and followed often equally by every other condition.

Balancing: A method of controlling extraneous variables by assuring that they affect the members of the experimental and control groups equally.

Basal age: In Stanford–Binet scale, the highest year level at which the subjects successfully passes all tests.

Basal level: A level for tests in which subtest items are ranked from easiest to toughest and below which the examinee would almost certainly answer all questions correctly.

Baseline: In small N design, the initial stage in which the behaviour to be changed is monitored to determine its normal rate of response.

Baseline stage: The first stage of a single N experiment in which a record of the individual's behaviour is made prior to any treatment or intervention.

Basement effect: A problem relating to measurement, whereby the researcher cannot measure the effects of an independent variable or its possible interaction because performance has reached minimum and is made prior to any treatment or intervention.

Basic research: A basic research is one in which abstract principles are identified for seeking knowledge about nature simply for the sake of understanding it better.

Between-group variance: A measure of variability among groups in an experiment.

Behaviour checklist: List of behaviours with predefined operational definitions that the investigators are trained for making use in any observational study.

Beta: A concept in signal-detection theory that refers to the observer's overall willingness to report that a signal has occurred.

Between-group design: An experimental design in which differen *t* groups of subjects serve under different treatment conditions of the study. Also known as between-subjects design.

Bias in construct validity: A type of bias that occurs when a test is shown to assess different hypothetical traits (psychological constructs) for one group than another or to assess the same trait but with different degree of accuracy.

Bias in content validity: A type of bias that occurs when an item or subtest is relatively more difficult for the individuals of one group than another, although the general ability level of the two groups is held constant.

Biased sample: A sample in which the major characteristics are systematically different from those of the parent population.

Bimodal distribution: A distribution of scores having two modes which need not have exactly the same frequency, and each clearly larger than any of others.

Bivariate analysis: A statistical analysis that investigates the relationship between two variables.

Bivariate prediction: A prediction of scores on one variable based on scores of one other variable.

Block randomization: A popular technique for carrying out random assignment in a random-groups design; here each block includes a random order of conditions and in each condition of the experiment, there are as many blocks as there are subjects.

Carryover effect: A relatively permanent effect that subjects in one condition have on their later behaviour in another condition.

Case study: Intensive investigation of a particular individual or case; it does not allow inferences of cause and effect and is merely descriptive in nature.

Category scaling: A technique of scaling in which the subject provides a rating for each stimulus on a scale with a fixed number of categories for indicating the absolute amount of quality being measured.

Ceiling effect: Occurs when scores on two (or more) conditions are at or near the maximum possible for the scale, giving the impression that no differences exist between these conditions.

Ceiling level: A level for tests in which subtest items are ranked from easiest to toughest and above which the examinee would almost certainly fail all remaining questions.

Centiles: Divisions that order scores by separating them into 100 equal parts.

Central tendency: Descriptive measure indicating the centre of a distribution of scores; includes mean, median and mode.

Central limit theorem: A mathematical principle that states that the distribution of sums (or means) of scores taken at random from any distribution of individuals will tend to form a normal or bell-shaped curve.

Checklist: A measure used to record the presence or absence of some behaviour in the situation under observation.

Chi-square test: A statistical test appropriate for analyzing nominal data usually obtained with between-group designs.

Classical theory of measurement: A dominant theory in psychological testing which assumes that an observed score consists of a true score plus error score or measurement score.

Closed questions: Questionnaire items in which response alternatives are provided.

Cluster sample: A type of probability sample in which the sampling unit is a cluster (or an aggregate) of elements for which a sampling frame is available.

Coding: An initial step in data collection in which units of behaviour or event are identified and classified according to specific criteria.

Coefficient of alienation (K): Measures the absence of relationship between X and Y and is calculated by $\sqrt{1-r^2}$.

Coefficient alpha: An index of reliability which is taken as the mean of all possible split-half coefficients, corrected by Spearman–Brown prophecy formula.

Coefficient of determination (r^2): Ratio of the explained variation to the total variation.

Coefficient of multiple determination (R^2): A measure of proportion of variation in the dependent variable that is explained jointly by two or more independent variables.

Cohort design: A design consisting of a longitudinal study of several groups.

Cohort effect: A situation which is produced when age differences are confounded by differences in subject history.

Complete counterbalancing: Occurs when all possible orders of conditions are used in a study based on within-subjects design.

Complete within-subjects design: A type of within-subjects design in which practice effects are controlled or balanced by administering the conditions several times to each subject so that the obtained results for each subject become interpretable.

Complex design: An experiment in which two or more independent variables are manipulated and studied simultaneously.

Componential Intelligence: The internal mental mechanisms in Sternberg's theory of intelligence that are responsible for intelligent behaviour.

Compromise design: See quasi-experimental design.

Concurrent validity: A type of criterion-related validity in which the criterion scores are obtained at approximately the same time as the test scores.

Confederate: A person offering his services to a researcher and who is instructed to behave in a certain way so that certain experimental treatment may be produced.

Confirmation: A process of subjecting a statement to empirical test.

Confirmatory experiment: An experiment in which a hypothesis is subjected to a test and confirmed.

Confounding: Simultaneous variation of an extraneous variable with an independent variable so that any effect on the dependent variable cannot be attributed with certainty to be the effect of the independent variable.

Constant: A term in a mathematical formula that does not vary with changes in conditions.

Constancy of condition: A technique of controlling extraneous variables by keeping them at a constant value for all conditions throughout the experiment.

Construct validity: A type of validity that indicates appropriateness of test-based interferences about the underlying construct supposedly measured by the test.

Content analysis: A technique for making inferences on the basis of objective and systematic identification of specific characteristics of messages.

Contextual intelligence: In Sternberg's theory, those mental activities which are involved in adaptation to real-world environments relevant to one's life.

Contingency table: A two-way table constructed for classifying data, with the major objective of determining whether the two directions of classifications are dependent (or contingent) upon one another.

Continuous variable: A variable that may assume any fraction and may change by any amount.

Control group: That group which does not receive the treatment in an experiment.

Criterion: A standard used to assess the validity of a test.

Control variable: A potential independent variable that is held constant in an investigation.

Convergent validity: A type of validity in which a test correlates highly with other variables or tests with which it shares an overlap of construct or with which it should correlate well.

Correction for guessing: Practice of revising a subject's final score in light of apparent guessing.

Correlation: A statistic showing the extent to which two variables are related, not necessarily in a casual relationship; the magnitude of a correlation can vary from –1.00 to +1.00.

Counterbalancing: A technique used to vary systematically the order of conditions in an experiment so that the distribution of the effects of the time of testing (that is, practice and fatigue) is not confounded with the conditions of the experiment.

Criterion-referenced test: A test where the purpose is to determine where the examinee stands with reference to a defined educational objective. No comparison is made to the performance of other examinees.

Critical value of a test statistic: Value or values that separate the rejection and acceptance regions in the statistical test.

Crossover interaction: A condition in which the effect of one independent variable on a dependent variable reverses at different levels of a second independent variable.

Cross-sectional studies: These are studies in which a large sample of population of various ages are taken at one time and tested (opposite of longitudinal studies).

Cross-sequential design: A research design that integrates both cross-sectional and longitudinal methods.

Crucial experiment: An experiment in which all possible hypotheses are simultaneously tested so that the true hypothesis may easily be ascertained.

Crystallized Intelligence: Includes abilities in Cattell and Horn's theory, which have been learned through the use of fluid intelligence.

Culture-fair test: A test which minimizes the irrelevant influences of cultural and social learning and thereby produces a separation of natural ability from specific learning.

Cumulative record: A record that displays the total number of responses during an experimental period as a function of time.

Concurrent validity: A type of criterion-related validity in which the criterion measures are obtained at approximately the same time as the test scores are obtained.

Confidence interval: A region of scores (that is, the scores between an upper and lower value) which is likely to include true population mean. Common confidence intervals are 95% and 99%.

Confidence limit: The two endpoints of a confidence interval.

Construct: A hypothetical state or factor (such as thirst) that can't be observed directly but is inferred from certain behaviour (drinking behaviour).

Contingency table: A two-dimensional table showing frequencies in each combination of categories of two nominal variables.

Convenience sample: A type of nonprobability sample in which the reseacher selects those persons from a group of people who meet the general requirements of the study.

Correlation for attenuation: Refers to correction of reduction caused by low reliability in the estimated correlation between a test and another measure. The formula for correction for attenuation is used to estimate what the correlation would be if the test and the criterion would have been perfectly reliable.

Correlation matrix: A table that tends to summarize a series of correlations among several variables.

Covariate: In ANCOVA, a variable that is controlled for or partialed out or held constant.

Criterion validity: A form of validity in which a psychological test is able to predict some future behaviour or is related meaningfully to some other measure.

Criterion variable: In regression analysis, the variable that is being predicted from the predictor variable.

Cross-lagged panel correlation: A type of correlational research for dealing with the directionality problem. Here if *X* and *Y* are measured at two different times and if *X* precedes *Y*, then *X* might cause *Y* and *Y* cannot cause *X*.

Cross-validation: In psychometrices the process of evaluating the validity of a test by administering it to an independent new sample drawn from the same population as the sample. Test validity is computed on a different sample of persons from that on which items were selected. In regression analysis, the measurement of predictor variables and comparison of predicted and observed values of the dependent variable on a fresh sample drawn from the same population as sample from which the regression equation was derived.

Curvilinear correlation: Any association between two variables other than a linear correlation.

d′: A statistic in signal-detection theory that relates to the sensitivity of the observer.

Data: The scores obtained on the measure of the dependent variable.

Data reduction: A process in analysis of behavioural data whereby the results are meaningfully organized and important statements about findings are summarized.

Debriefing: A procedure by which the researchers inform subjects about all aspects of a study or research after they have participated in it so that any negative consequences of the procedure may be removed.

Deception: A strategy in research in which the participants are not told of all the details of the experiment at its outset; generally used for the purpose of avoiding demand characteristics.

Deciles: Scales that divide the frequency distribution into equal tenths.

Degree of freedom: The number of values that are free to vary, assuming that the sum of values and the number of values are fixed.

Dehoaxing: That part of the debriefing in which the true purpose of the study is explained to the participants.

Demand characteristics: Cues within the research context or the situations of the experiment that guide the subject's behaviour.

Dependent sample: Two or more groups for which the selection of the members of one group tends to determine the characters of the members of the other groups.

Dependent variable: The variable measured and recorded by the experimenter in order to assess the effect of independent variables; it is hypothesized to be dependent on the value assumed by the independent variables. The dependent variable is also dependent upon extraneous variables, hence requires control procedures.

Descriptive statistics: Statistical methods popularly used in organizing and summarizing data.

Design of the experiment: A specific systematic plan for varying independent variables and noting the likely changes in the dependent variables.

Dichotomous format: A test item format in which there are two alternatives for each item.

Dichotomous scale: A nominal scale that consists of only two categories.

Difference threshold: The increment or decrement in the differences of the two stimuli values which can be detected 50 per cent of the time by an observer.

Differential transfer: A problem in a within-subjects design when performance in one condition differs depending upon which of the two conditions comes first.

Differential validity: The extent to which a test has different meanings for different groups of people. For example, a test may be a valid predictor of success in the concerned field for tribal people but not for nontribal people.

Direct replication: A process of repeating an experiment as closely as possible to test whether or not the same results will be obtained.

Directionality problem: In correlation research directionality problem refers to the fact that for a correlation between *X* and *Y*, it is possible that *X* is causing *Y* but it is also likely that *Y* is causing *X*. The correlation also does not provide any basis for deciding between these two alternatives.

Discordant pair: Two cases are ranked on the opposite order on two variables.

Discrete variable: A variable in which each level represents a distinct category that is qualitatively different from another category—for example, male and female.

Discriminant validity: A type of validity which results when a test does not correlate with variables or tests from which it differs.

Dispersion: The extent of the spread of scatteredness in a distribution of scores.

Distractor: An alternative on a multiple-choice item that is apparently similar but not correct and no credit is given for the same.

Distribution-free statistics: Statistical tests that make no assumptions about the form of the distribution of the sample.

Divergent thinking: A kind of thinking that goes off in different directions.

Double-blind technique: A technique in which both the subject and the observer, or the experimenter, are kept unaware (blind) of what treatment is being administered.

Ecological validity: The extent to which an experimental situation can be generalized to behaviours occurring in natural settings or day-to-day living.

Effect size: Standardized measure of difference between population means. Effect size increases with great differences between means and tends to decrease with greater standard deviations but is not affected by same size.

Element: Each number of the population of interest.

Elimination: A technique for controlling the effect of extraneous variables by removing them from the experimental situation.

Empirical approach: An approach which emphasizes the acquisition of knowledge from direct observation and experimentation for answering questions.

Empiricism: A way of getting knowledge that relies on direct observation or experience.

Equal-interval variable: Those variables in which the numbers stand for about equal amounts of what is being measured.

Error of estimation: Distance between an estimate and the true value of the parameter.

Error variance: The sum of all uncontrolled sources of variation that are likely to affect the dependent variable: it is understood as the denominator of the t or F test.

Error variation: Nonsystematic variation in an experiment; usually estimated on the basis of differences among subjects within each group in a random-group design.

Evaluation apprehension: A form of anxiety by the participants that leads them to behave so as to be evaluated positively by the experimenter.

Evidence report: The summarized results of an empirical study which are commonly used in confirmation of a hypothesis.

Existential hypothesis: A conjectural statement which asserts that a relationship holds for at least one case.

Expectancy effect: Tendency for the results of the experiment to be influenced by what the experimenter expects to find. Also known as Rosenthal effect, named after a psychologist who has studied this problem extensively.

Expectancy table: A table that shows the established relationship between test scores and expected outcome on a relevant task.

Experimental control: Conditions in experiment in which extraneous variables are held constant so that any effect upon the dependent variable can be easily attributed to the manipulation of the independent variable.

Experimental design: A design for obtaining and treating data in which the experimental method is used.

Experimental error: See Error Variance.

Experiential intelligence: In Sternberg's theory, the ability to deal effectively with novel tasks.

Experimental method: A method in which one or more independent variables are manipulated and responses on one or more dependent variables are measured.

Experimental realism: The extent to which subjects are psychologically engaged in an experimental task such that they become less concerned with demand characteristics.

Experimental units: Objects upon which measurements are taken.

Experimenter bias: An effect that an experimenter may unknowingly exert on the results of an experiment.

Ex post facto: Refers to conditions in an experiment where some manipulation has occurred naturally prior to the start of the experiment (literally it means from after the fact).

Expected frequency: In chi-square test, the number of people in a category or cell expected if the null hypothesis is true.

External validity: Extent to which the results of a study or a research can be generalized to different populations or settings.

Extraneous variables: Variables which may potentially influence the results of a study but are not the variables whose effects are to be studied by the experimenter.

Extravalidity concerns: Side effects and unintended consequences of testing.

***F* test:** A ratio between larger variance and smaller variance.

Face validity: Not an evidence for validity; the extent to which items on a test appears to be meaningful and relevant.

Factor: Independent experimental variables which are related to a dependent or response variable. In factor analysis, a group of variables that correlate maximally with each other and minimally with variable not in the group.

Factor analysis: A set of multivariable data analysis methods for reducing larger matrices of correlations to fewer variables.

Factorial design: A complex experimental design in which all possible combinations of the selected values of each independent variable are used.

Factor loading: In factor analysis, correlation of a variable with a factor.

Factor matrix: A table of correlation between variables and factors. These correlations are called as factor loadings.

False alarm: Incorrectly telling about the presence of a signal on a trial where only noise occurred.

False negative: In test-decision analysis, a case in which the test suggests a negative classification, yet the correct classification is positive; the subject is incorrectly predicted to fail on the criterion.

False positive: In test decision analysis, a case in which the test suggests a positive classification, yet the correct classification is negative; the subject is incorrectly predicted to succeed on the criterion.

Field experiment: An experiment in which the experimenter manipulates one or more independent variables in their natural settings to see the impact upon the dependent variables.

Field survey: A method of research in which subjects complete a questionnaire or interview in a natural setting.

Filter questions: General questions framed in a survey to determine whether the respondent needs to be asked some specific questions later on.

Fixed-effects model: A statistical model in which all values of the independent variables about which inferences are to be drawn are specified in the design.

Floor effects: An effect that results when a task is too difficult, causing most or all scores to the lowest possible score.

Fluid intelligence: In Cattell & Horn's theory, a largely nonverbal and culture-reduced form of mental efficiency.

Forced-choice method: An item-writing method in which alternatives are matched for social disability and the raters are forced to choose between options that are equal in social desirability.

Frequency: Number of scores or observations falling in a cell or classification category.

Frequency distribution: A particular distribution in which a set of scores are arranged in ascending or descending order; the number of times each score occurs is also indicated.

Friedman test: A nonparametric statistic appropriate for analyzing ordinal data obtained with a within-subjects design.

Fundamental lexical hypothesis: In personality theory, the idea that trait terms have survived in language because they have conveyed important information about our dealing with others.

Funnel questions: Questions dealing with a particular topic in a questionnaire, ordered from the most general to the most specific.

G factor: In Spearman's theory, a general factor of intelligence that must exist to account for the observed correlations between several tests; a general mental ability that is thought as a kind of mental energy.

Generality of results: Whether or not a particular experimental result will be obtained under different settings, such as using subjects from a different population.

Good subject role: A form of participant bias in which participants try to guess the experimenter's hypothesis (or mindset) and then behave in a such a way as to confirm it.

Grade norm: A type of norm that displays the level of test performance for each grade in the normative sample

Grand Mean (GM): In ANOVA, the overall mean of all scores regardless of what group they are in.

Group test: Mainly paper-and-pencil test suitable for testing large groups of persons at the same time.

Halo effect: Rater's tendency to rate a person high or low on all dimensions because of a global impression.

Hawthorne effect: Condition where performance in an experiment is affected because the subjects know that they are being observed.

Heavy-tailed distribution: A distribution that differs from a normal curve by being too spread out so that the histogram of the distribution would tend to have too many scores at each of the two tails.

Heritability Index: A broad estimate that shows how much of the total variance in a given trait is due to genetic factors. This index may vary from 0.0 to 2.0.

High-order interaction: An interaction effect involving more than two independent variables in experiments.

Histogram: A graphical method for describing a frequency distribution in which the height of the bars indicates the frequency of class intervals.

History: A threat to internal validity that occurs when some historical events affect the participants between the beginning of a study and its end.

Hit: A correct detection of signal presented to the observer.

Homogeneity of variance: A statistical assumption showing equal variances among groups under study.

Homogeneous test: A test in which the individual items tend to measure the same thing; this is assessed by item-total correlations.

Homoscedasticity: A statistical assumption which asserts that the scatter of X values around Y values is approximately equal.

Hypothesis: A testable statement of the relationship between variables that is advanced as a tentative solution to a problem.

Incomplete within-subjects design: A type of within-subjects design in which each condition is presented to each subject only once, and the order of presenting the various conditions is varied across subjects such that practice effects are neutralized by pulling together the results of all subjects.

Independent sample: Two or more groups for which the relation of the members of one group does not have an impact upon the membership of the other group.

Independent variable: The variable manipulated by the experimenter for the purpose of determining whether it influences behaviour.

Infactor analysis: A group of variables that correlate maximally with each other and minimally with variable not in the group.

Inferential statistics: Statistical measures for testing whether differences in the dependent variables (or responses) that are commonly associated with various conditions of an experiment are reliable; such statistics are employed to make inferences about population.

Informed consent: The notion that persons should be given sufficient information about a study so that their decision to participate as a research subject remain an informed one.

Instructional variable: Type of independent variable in which participants are given different sets of instructions about how to perform.

Intelligence quotient: It is a ratio of the individual's mental age (MA) as determined by the test to actual or chronological age (CA). Thus IQ = MA/CA 100.

Interaction: A condition in which the effect of one independent variable differs depending upon the level of the second independent variable.

Intercept: On a two-dimensional graph, that point on *Y* axis where *X* equal zero. In regression, this is the point at which the regression line intersects the *Y* axis.

Internal consistency reliability: The extent to which items of a test assess a common characteristic, usually measured by split-half reliability and Cronbach's alpha.

Internal validity: The extent to which differences in dependent variables can be clearly attributed to an effect of independent variables as opposed to the effect of extraneous variables. In other words, the extent to which any experiment or research study is methodologically sound and free from confounding.

Interobserver reliability: The extent to which two independent observers are in agreement with each other.

Interrupted time series design: A form of quasi-experimental design in which a treatment is evaluated by measuring performance several times prior to the institution of treatment and several times after the treatment has been given.

Interval estimation: The determination of an interval within which the population parameter is presumed to lie. Confidence interval is an example of an interval estimation.

Interval scale: A measurement scale that can be used to rank objects and on which the units reflect equivalent magnitudes of the property being measured.

Interviewer bias: A kind of bias shown by the interviewer in which he tries to adjust the wording of a question to suit the respondent or tries to record only the favourable position of the respondent's answers.

Ipsative score: Such scores compare the individual against him or herself and thus each person provides his or her own frame of reference.

Ipsative test: A test in which the average of subscales is always the same for every examinee; for this, an individual examinee's high score on subscales must be balanced by low scores on other subscales.

Item analysis: A set of methods used to assess item difficulty and item discriminability.

Item characteristic curve (ICC): A curve prepared for each item and shows the total test score on the *X*-axis and proportion of test takers passing the item on the Y-axis.

Item difficulty: An index of item analysis used to assess how difficult items are, the proportion passing an item.

Item difficulty index: Proportion or percentage of examinees who get a particular item correct.

Item discrimination: An index of item analysis used to assess how well item separate groups who score high and low on the test.

Item discrimination index: A statistical index that shows how efficiently an item discriminates between persons who obtain high and low scores on the entire test.

Item reliability index: The product of a test item's internal consistency as indexed by the correlation with the total score and its variability as indexed by the standard deviation.

Item response theory (IRT): A modern theory of test construction in which the psychometricians posits a single dimension of skill or underlying trait on which all test items rely; it is also known as latent trait theory because here each examinee is assumed to have a certain amount of the latent trait being measured. The theory rests on the assumption that the probability of a particular response to a test item is the product of one or more characteristics of the individual and one or more characteristics of the test item.

Item-validity index: The product of a test item's standard deviation and the point-biserial correlation with the criterion.

Just-noticeable difference: See difference threshold.

Kuder–Richardson formula 20: A formula for estimating the internal consistency reliability of a test. For use of this formula, all items must be scored either 0 or 1.

Kurtosis: Extent to which distribution of scores deviates from a normal curve having tails that are too thick or too thin.

Laboratory research: A research that is conducted within controlled conditions of the scientific laboratory.

Latin square design: A design in which form of partial counterbalancing occurs where each condition of the study occurs equally, often in each sequential position, and each condition precedes and follows each other condition exactly one time.

Law of large numbers: As the size of the sample increases, the standard error of mean tends to decrease.

Level of significance: Level against which the probability of an outcome under the null hypothesis is held to be statistically significant or not significant; also known as alpha (α) level.

Light-tailed distribution: A type of distribution that tends to differ from a normal distribution curve by being too peaked or pinched so that a histogram of the distribution would have too few scores at each of the two tails.

Likert scale: A scale constructed by Likert which presents the examinee with five responses ordered on Agree/Disagree continuum.

Linear relationship: Such relationship between two variables in which an increase in one variable is always followed by a constant increase or constant decrease in the other variable.

Local norms: Norms based upon a representative local sample as opposed to a national sample.

Longitudinal study: A method of study in which individuals are studied overtime, and measurements are taken on the individuals at various time intervals.

Main effect: The overall effect of an independent variable in a factorial design.

Mann–Whitney *U* test: A nonparametric statistic for comparing two populations based on independent random samples from each.

Marginal mean: In factorial design in ANOVA, mean score for all participants at a particular level of one of the variables.

Matched-groups design: A type of design in which the experimenter forms comparable groups by matching subjects on a pretest task and subsequently, randomly assigning the members of these matched sets to the different conditions of the experiment.

McCall's *T*: A standardized score system having a mean of 50 and standard deviation of 10. It can be easily obtained from a simple liner transformation of z score, that is, $T = 10z + 50$.

Mean: A measure of central tendency; the sum of all scores divided by their numbers; more popularly known as arithmetic mean.

Mean square: Sum of squared deviations from the mean divided by degree of freedom. Also known as *variance estimate*.

Measurement: A process of assigning numerals to objects or events according to some rules.

Measurement error: Everything other than the true score that makes up an examinee's obtained test score.

Mechanical subject loss: A condition which occurs when a subject fails to complete the experiment either due to equipment failure or due to some errors made by the experimenter.

Median test: A statistical tool appropriate for analyzing ordinal data obtained for a study based upon a between-subjects design.

Mental age: A unit for expressing the results of intelligence tests.

Mesokurtic: A distribution which is neither peaked nor flat and, therefore, normal. So mesokurtic is normal distribution.

Meta analysis: Statistical procedures for combining, and describing the results from different studies.

Method of empirical keying: A method of scale development in which test items are selected based entirely on how well they contrast with a criterion group taken from a normative sample.

Method of rational scaling: A scale construction method in which all items of scale correlate positively with each other and also with the total score for the scale. Also known as internal consistency approach.

Mixed design: A complex design in which different designs are used for the different independent variables in the experiment.

Mode: A measure of central tendency; that score which occurs most frequently in the distribution.

Monotonic relationship: A relationship between two variables in which an increase in one variable is always accompanied by a consistent increase or decrease in the other variable.

Mutiple-baseline design: A single *N* experimental design in which the effect of a treatment is displayed by demonstrating that behaviours in more than one baseline change as a consequence of the introduction of a treatment; in fact, multiple-baselines are established for different individuals, for different behaviours in the same individual in different situations.

Multiple correlation coefficient (R): A measure of linear correlation between a criterion or dependent variable and two or more predictor or independent variables.

Multitrait-Multimethod matrix: A type of research design for assessing convergent and discriminant validity in such a way that calls for the assessment of two or more traits by two or more methods.

Multivariable analysis: A set of methods for data analysis that considers the relationship between combination of three or more variables.

Multivariate analysis of covariance (MANCOVA): Analysis of covariance with more than one dependent variable.

Multivariate analysis of variance (MANOVA): Analysis of variance with more than one dependent variable.

Mundane realism: The extent to which the experiment mirrors real-life experiences. Generally considered less important than experimental realism.

Natural-groups design: A type of design in which the conditions tend to represent the selected levels of a naturally occurring independent variable like age variable (of the subject).

Naturalistic observation: A type of observation made in more or less natural settings without any kind of intervention by the observer.

Negative correlation: A particular type of relationship between two variables in which high scores on one go with low scores on the other, lows with highs and mediums with mediums.

Nominal measurement: A measurement by which the subjects are placed into qualitatively different categories.

Nondirectional hypothesis: A research hypothesis that does not predict a particular direction of difference between populations.

Nonlinear effect: Any outcome that does not form a straight line when plotted on a graph. It occurs only when the independent variable has more than two levels.

Nonparametric statistics: See distribution-free statistics.

Nonparametric statistical test: A statistical test that makes no assumptions about population parameters; also known as distribution-free test.

Nonprobability sampling: A type of sampling procedure in which there is no way to estimate the probability of each element in the population being included in the sample.

Nonreactive measure: A measure of behaviour in which neither the subject is aware of being observed nor is his behaviour changed by the process of observation.

Normal distribution: A symmetric, mesokurtic, bell-shaped form of distribution of scores.

Normalized standard score: A score that is obtained by a transformation that renders a skewed distribution into a normal distribution.

Normative sample: A group consisting of individuals who have been administered a test with the standard instructions, format and general procedures outlined in the manual for administering the test. Also called as standardization sample.

Norm-referenced test: A test that assesses each person relative to a normative group.

Norms: An average performance of the standardization sample on the test.

Null hypothesis: An assumption used as the first step in statistical inference, whereby the independent variable is said to have no effect upon the dependent variable which is not influenced by the manipulation in the independent variable.

Observer bias: Systematic errors occurring in observations due to the observer's expectations regarding the outcome of a study.

One-shot, or cross-sectional study: A survey design in which one or more samples of the population are taken and information is collected from the samples at one time.

One-tailed test: A statistical procedure for testing the null hypothesis in which the entire rejection area is kept at one end of the sampling distribution.

Open-ended question: A question that usually cannot be answered specifically and requires the examinee to produce something spontaneously.

Operant level: See baseline stage.

Operational definition: A definition of a concept in terms of operations that must be completed for defining that concept, that is, definition of the concept in terms of which it is measured. For example, intelligence is what the intelligence test measures.

Order effects: The influence on a particular trial that arises from its position in the sequence of trials.

Ordinal scale: A scale that orders or ranks objects or events from most to least, or least to most without providing information on the exact distance between scale points or categories.

Ordinate: The vertical axis (Y-axis) in a graph.

Outlier: Score with an extreme (very high or very low) value in relation to other scores in the distribution.

Parallel forms: Two alternative forms of a test having equal means, equal variances and equal inter-item correlation.

Parameter: A measure obtained from all possible observations in a population.

Partial correlation: A technique of correlation in which two variables are correlated while the effect of the third variable is held constant.

Partial counterbalancing: Occurs when subset of all possible orders of condition is used in within-subjects design.

Partial replication: Repetition of a portion of some prior research.

Participant bias: Occurs when the behaviours of the subjects or participants is influenced by their beliefs about how they are supposed to behave in a study.

Percentile: Percentages of the individuals in the standardization sample who scored below a specific raw score. It varies from 0 to 100.

Percentile band: The range of percentiles that are likely to represent a subject's true score. It is developed by forming an interval one standard error of measurement above and below the obtained score and converting the resulting values into percentiles.

Personality coefficient: A term used by Mischel to refer to the finding that the predictive validity of personality scales rarely exceeds 0.30; an attack on trait concept of personality.

Percentile rank: Proportion or percentage of scores that fall below a particular score.

Phenomenological report: Subject's description of his or her own behaviour or conditions of mind in a study; also called subjective report.

Pilot study: During initial stages of research, some data are collected and the problems spotted in these initial stages help the researchers in refining the procedures and prevent the full-scale study from being flawed from the point view of methodology.

Placebo effect: Improvement in performance regardless of the nature of the treatment; this is because here a suggestion is made that participation in the research study is likely to result in something positive happening to the participant.

Plagiarism: Taking the ideas of others deliberately and claiming them as one's own.

Point of subjective equality (PSE): That point which is perceived as psychologically equal to a specified standard on a physical scale.

Point scale: A scale or test in which points such as 0, 1 or 2 are assigned to each item.

Polytomous format: Also known as polychotomous format, a format for objective test in which three or more alternative responses are given to each item.

Population: A well-defined group of people, object, etc.

Positive correlation: A relationship between two variables in which high on one is associated with high on other and low on other is associated with lows on other.

Post hoc fallacy: Faulty causal inferences from the correlational data.

Power test: A test that allows enough time for the examinees to attempt all items.

Practice effects: The influence on performance which arises from practicing a task.

Predictive validity: A type of criterion-related validity in which the criterion measures are obtained in future, usually months or years after the test scores are obtained.

Predictor variable: In regression analysis, the variable used to predict the criterion variable.

Pretest-Post-test design: A popular research design in which subjects are measured before and after a treatment.

Primary mental abilities: A group of seven factors of intelligence posited by Thurstone.

Probability: An estimate of the likelihood that a particular thing or event will occur.

Probability sampling: A sampling technique in which the probability of an element being included in the sample is specified.

Problems: Research questions that initiate scientific inquiries.

Progressive effect: Any sequential effect in within-subjects design in which the accumulated effects are assumed to be the same from trial to trial (such as fatigue).

Projective hypothesis: A hypothesis which states that the personal responses towards the ambiguous stimuli reflect the unconscious needs, motives and conflicts of the individual.

Projective test: A test in which the examinee is given a vague, ambiguous stimuli and responds with his or her own ideas or constructions.

Prophecy formula: A formula developed by Spearman and Brown that is used for correcting the loss of reliability that occurs when the split-half method is used. The method is also used to estimate how much the test length must be increased to bring the test to a derived level of reliability.

Pseudo problems: Research questions that may appear to be real questions but on further examinations, it becomes clear that they can't be attacked through empirical means.

Psychological abstracts: A monthly publication that summarizes the studies recently published in different journals of psychology.

Psychological testing: It refers to all the possible uses, applications and underlying concepts of psychological tests.

Psychophysics: A study of quantitative relationship between the physical stimulus and the resulting psychological sensation.

Purposive manipulation: Systematic control of an independent variable during experimentation.

Purposive sample: A type of nonprobability sample in which the elements to be included in the sample are selected by the researcher on the basis of special characteristics or typicalities of the respondents.

***P* × *E* factorial design:** A factorial design in which at least one subject (*P* means person variable) and one manipulated variable (E means environmental variable) are found.

***Q* technique**: A technique for studying changes in self-concept or other variables by sorting statements into assigned categories.

Qualitative factor: Factors that are not quantitative.

Quantitative factor: Factors that may take on different values corresponding to the points on a real line.

Quartile deviation (Q): The average of the differences between the 3rd and 1st quartile; divides the frequency distribution into equal fourths.

Quasi-experimental design: A research design that resembles a true experimental design but subjects are not randomly assigned to treatments nor are the treatments randomly assigned to the group; such a design is basically an attempt to stimulate the true experiment whose criteria are only partially met, that is, control over manipulation of independent variables is somehow achieved but control over assignment of subjects into equivalent groups is not achieved.

Random assignment: A procedure adopted in placing the subjects in groups or to order events such that only chance factors may determine the placement of ordering.

Random-groups design: A type of between-subjects design in which random assignment of subjects to different conditions is done.

Random sampling: A procedure of sampling in which each element of the population has an equal chance of being selected in the sample, and also selection or nonselection of one subject cannot influence the selection of the other.

Range: A measure of dispersion; a difference score obtained by subtracting the smallest score from the largest score in the distribution.

Rank-order scaling: A procedure of ranking in which the subjects are presented all the stimuli or objects simultaneously with a request to rank them from highest to lowest or the vice versa.

Rater bias: A tendency for rater's ratings to be inaccurate because of leniency, severity and other types of evaluation errors.

Ratio scale: A scale in which there are equal intervals between scale values and there is an absolute zero.

Raw score: An elementary and basic level of information provided by a psychological test, for example, the number of questions answered correctly.

Rective measure: Measure of behaviour under circumstances in which the participants are aware that their behaviours are being observed or there is reason to believe that the observation or measurement procedures will influence the participant's behaviour.

Reactivity: A special phenomenon that causes the reliability of scale to be higher when an observer knows that his or her work is being monitored.

Regression equation: An equation that describes the best-fitting straight line for estimating the criterion from the test.

Relevant variable: A variable that has been shown to influence the behaviour (or dependent variable) directly or indirectly.

Reliability: The extent to which the same observations or scores are obtained in repeated studies; usually computed by a correlation between scores obtained by the same subjects on two forms of the test or scores on the same test at two different time intervals or scores obtained on each half of the test.

Reliability coefficient: Ratio of true score variance to total variance of the test scores.

Repeated-measures design: A design in which participants are tested in each of the experimenter's conditions. Also known as within-subjects design.

Replication: Repeating an experiment that occurs primarily when the results of some prior study are suspected to be erroneous.

Representativeness: A sample is said to be representative if it has the same distribution of major characteristics as the population from which it was selected.

Representative sample: A sample consisting of individuals with characteristics similar to those for whom the test is ultimately needed.

Research hypothesis: The assertion that an independent variable will have an effect upon a dependent variable.

Residual: The difference between predicted and observed values from a regression equation.

Respondent mortality: Condition that results when a respondent fails to complete all phases of a study or longitudinal research.

Response bias: Any systematic tendency of the respondent to choose a particular response category in a questionnaire for an extraneous reason, not related to variable that the response is supposed to indicate but related to the content of the question.

Response style: A tendency to mark a test item in a certain way, irrespective of meaning or content of it.

Reversal design: A research design in which the experimenter alternates between the baseline condition and the treatment condition.

Reverse counterbalancing: Occurs in a within-subects design when the subjects or participants are tested more than once per condition; they are given one sequence, then a second with the order reversed from the first such as ABBA or ABCCBA.

Robust test: A statistical test from which statistical inferences are likely to be valid even when there are departures from normality in the population distribution.

Rosenthal effect: See expectancy effect.

Routing procedure: A procedure in which first items or subtests are administered for the purpose of determining the appropriate starting points for subsequent subtests. This procedure has been included in Stanford–Binet fifth revision.

Sample: A group of persons selected from the population.

Sampling distribution: A theoretical probability distribution of a statistic that would result from drawing all possible samples of a given size from the same population.

Savant: A mentally deficient person but who shows a highly developed talent in a single area such as art, memory, music, etc.

Scale: A set of numerals assigned to objects or events indicating the relative amounts of some characteristics that are possessed by those objects or events.

Scaling techniques: Scientific procedures for measuring stimuli on psychological dimension.

Scatter-plot (Scatter diagram): A group of scores made by the same individuals on two different variables, providing a pictorial representation of the degree of relationship between these two variables.

Self-report questionnaire: A questionnaire that provides statements about an individual who is required to answer each statement as 'True' or 'False'.

Sematic differential scale: A rating technique in which the rater uses a seven-point scale to rate a concept on a number of bipolar adjectives such as good-bad, strong-weak, active-passive.

Sequence effect: Occurs in a within-subjects design when experience of participating in one of the conditions of the study influences performance in subsequent or later conditions.

Serendipity: An unexpected but more valuable finding than the original purpose of the research.

Shrinkage: Sometimes a regression equation, created for one group, is used to predict the performance of another group of participants. This procedure tends to overestimate the magnitude of the relationship for the second group. The amount of decrease in the strength of the relationship from the original sample to the sample with which the equation is used, is technically known as shrinkage.

Significance: Significance has two basic dimensions—statistical significance and psychological significance. Statistical significance indicates whether the obtained results are a common or rate an event if only chance is operating. Psychological significance indicates qualities of data, adequacy of the data obtained and the clarity of the obtained results.

Significance level: See level of significance.

Simple main effect: The effect of one independent variable at one level of a second independent variable in a factorial design.

Single-factor design: Any experimental design with a single independent variable.

Single-factor multilevel design: Any design with a single independent variable and more than two levels of the independent variable.

Single-*N* experimental design (or Single-subject experimental design): An experimental procedure that focuses on behaviour changes in an individual by systematically observing and monitoring the individual's behaviour.

Situation sampling: The random or systematic selection of the situations in which observations are to be made for the purpose of ensuring the goal of representativeness.

Situational variable: A type of independent variable in which subjects encounter different environmental situation or circumstances.

Skewed distribution: A nonsymmetrical distribution; in a negatively skewed distribution, extreme scores are below the mean of the distribution and in a positively skewed distribution, extreme scores are above the mean of the distribution.

Skewness: Asymmetry of a frequency distribution; positive skewness shows that scores are piled up at the low end and negative skewness shows that the scores are piled up at the high end.

Social desirability: A condition in which the respondent tends to answer in terms of what is socially most desirable or acceptable rather than in terms of what they believe to be true.

Social desirability response set: Tendency on the part of the examinee to respond to test items in the perceived desirable manner.

Social intelligence: A mental ability to understand other people and to relate effectively to them.

Source traits: Stable and constant traits that control behaviour; less visible than surface traits but more important in accounting for behaviour.

Spearman–Brown prophecy formula: A formula for adjusting split-half correlations so that they may point out the full length of a scale.

Speed test: A test which has strict time limit, enough that few subjects will finish answering of all items; it contains items of more or less uniform difficulty.

Split-ballot technique: A technique used in questionnaire construction in which different wordings of the same questions are used for two (or more) equivalent groups of respondents to examine the effect of wording directly.

Split-half reliability: A method of estimating reliability in which test items are divided into two equal halves and subsequently, scores obtained on these two halves are correlated.

Spurious correlation: When evidence falsely indicates that X variable and Y variable (sometimes more than two variables) are associated.

Spurious relationship: Occurs when the zero-order relationship between an independent and dependent variable disappears or becomes significantly weaker with the introduction of a control variable.

Standard deviation: A measure of dispersion; square root of the sum of squared deviations of each score from the mean divided by the number of scores.

Standard error: Standard deviation of the sampling distribution.

Standard error of estimate: A statistical index for the accuracy of a regression equation. It is defined as the standard deviation of the residuals from a regression analysis. When standard error of estimate is small, prediction is considered most accurate.

Standard error of measurement: An index of measurement error that indicates the extent to which an examinee's score might vary over a number of parallel tests.

Standard score: A derived score in which the original score is expressed as the distance from the mean in standard deviation units; the standard about the standard score is that it has a fixed mean and a fixed standard deviation.

Standardization fallacy: A fallacious view that a test is standardized on one sample is ipso facto unfair when used in other population.

Stanine system: A system for assigning the numbers 1 through 9 to a test score.

Statistic: A numerical value that is computed from the observations of a sample taken from population.

Statistical determinism: An assumption made by research psychologist that behavioural events can be predicted with a probability greater than chance.

Sten Scale: A 10-unit scale with five units above and five units below the mean.

Stratified random sampling: A type of probability sampling in which a population is divided into various strata and random samples are taken from each of these strata.

Structured Interview: An interview conducted under well-defined procedure in which questions or sequence of questions are predefined and predetermined.

Subgroups norms: Norms derived from an identified subgroup as opposed to a national sample that is diversified.

Subject: The organism or object on which observation or manipulation is done.

Subject variable: A type of independent variable that is selected rather than manipulated by the researcher; it refers to already existing attribute of the individual chosen for the investigation such as age, gender.

Successive independent samples: A research design in which several cross-sectional surveys are done and the same questions are asked from each succeeding sample of participants or respondents.

Surface traits: Traits which are readily observable and are most obvious aspects of personality.

Systematic measurement error: A type of measurement error that arises when, without the knowledge of test developer, a test consistently measures something other than the trait, which it intended to assess.

Systematic observation: Observations done according to some explicit procedures, as well as in accordance with the logic of scientific inference.

Systematic variance: Variability that can be attributed to some identifiable source, either systematic variation of the independent variable or the uncontrolled variation of a confounding variable.

Task variable: A type of independent variable in which participants are given different types of tasks to perform.

Temporal validity: The extent to which the obtained experimental results can be generalized to other time frames.

Test battery: A group of tests, the scores of which are used together in assessing an individual.

Testable statement: A statement that may be directly or indirectly subjected to empirical verification.

Test-retest reliability: A method of estimating reliability in which the same test is given twice to the same group of heterogeneous subjects and the resulting scores are correlated.

Theory: A statement of relationship between two or more than two variables.

Third variable: A variable that may account for the observed relationship between two other variables.

Threats to validity: Possible factors or causes which must be controlled so that a clear cause-and-effect inference can be drawn.

Tied pair: Two cases are ranked similarly on one or both of two variables.

Time sampling: Systematic or random selection of observation intervals for ensuring a representative sample of behaviour.

***t*-ratio:** A test statistic for determining the significance of a difference between two means.

True score: An individual's hypothetical real score on a test; it can be determined on probability basis and is never directly known.

***T*-score:** A standard score with a mean of 50 and standard deviation of 10.

Two-tailed test: A statistical procedure for testing the nondirectional null hypothesis in which the rejection areas are placed at both the ends of the sampling distribution.

Type I error: An error in which the null hypothesis is rejected even when it is true.

Type II error: An error in which the null hypothesis is accepted even when it is false.

Unimodal distribution: Frequency distribution with one value clearly having the larger frequency than any other.

Universal hypothesis: A hypothesis that asserts that the relationship for the given variables holds for all times and at all places.

Unobtrusive measure: Any measure of behaviours that can be recorded without participants knowledge that their behaviour has been observed.

Unstructured Interview: An interview conducted without any specific or particular questions or sequences of questions.

Validity: The extent to which the test measures what it intends to measure.

Validity shrinkage: A discovery in cross-validation research that a test predicts the relevant criterion less accurately with the new sample of examinees than with the original tryout sample.

Value: A shared and enduring belief about ideal modes of behaviour or end states of existence.

Variable: Anything that can assume different numerical values and can be measured or manipulated.

Variance: The square of the standard deviation; a measure of the average squared deviation of a set of scores from the mean score.

Verification: A process of determining whether a hypothesis is probably true or false.

Volunteer bias: The bias that results from the fact that a given sample contains only those subjects who have shown their willingness to participate in the study.

Weber's law: The just-noticeable difference between two stimuli which can be stated as the ratio between the two stimuli that are independent of their magnitude.

Weighted mean: The sum of the mean of each group multiplied by its respective weight (the *N* in each group), divided by sum of the weights (total *N*).

Withdrawal design: Any small N design in which a treatment is in place for a time and is then removed to determine if the rate of behaviour comes to the baseline.

Within-group variance: A measure of dispersion or variability among subjects in the same group.

Within-subjects design: An experimental design in which each subject is tested under more than one level of independent variable.

Written survey: A survey method in which the investigator prepares a written questionnaire that is filled out by the participants.

***y*-axis:** The vertical axis of the graph.

***y* prime:** The value of y that falls on the regression line above only x. It is symbolized as y.

Yoked control group: A control group in which the treatment given to a member of the control group is matched exactly with the treatment given to a member of the experimental group.

***Z*-score:** A standard score showing distance of each score from the mean of the distribution in terms of standard-deviation unit.

***Z*-test:** Hypothesis testing procedure in which there is a single sample and the population variance is known.

◆